3

POSITIONNEMENTS

Toute activité de recherche serait, selon certains acteurs de celle-ci, nécessairement politisée, de par pour commencer le choix de ses objets. Ainsi, Ripoll alerte contre l'illusion d'une recherche objective et les dangers de la technocratie [Ripoll, 2017]. Nous ne rentrerons pas dans ces débats bien trop vastes pour être traités même en un chapitre, puisqu'il rejoignent des thèmes de sciences politiques, d'éthique, de philosophie, liés par exemple à la gouvernance scientifique, à l'insertion de la science dans la société, à la responsabilité scientifique.

Il est clair que même des sujets a priori intrinsèquement objectifs, comme la physique des particules et des hautes énergies, ont des implications regardant d'une part les choix de leur financements et les externalités associées (par exemple, l'existence du CERN a largement contribué au développement du calcul distribué), mais d'autre part aussi les applications potentielles des découvertes qui peuvent avoir des répercussions sociales considérables. En biologie, l'éthique est au coeur des principes fondateurs des disciplines, comme en témoignent les débats soulevés par l'émergence de la biologie synthétique [Gutmann, 2011]. Les tenants d'approche prudentes dans celle-ci se recoupent avec la biologie intégrative, or les sciences intégratives défendues par Paul Bourgine, mises en oeuvre par l'intermédiaire du campus digital Unesco CS-DC[1], ont typiquement la responsabilité sociale et l'implication citoyenne au coeur de leur cercle vertueux. En sciences humaines et sociales, comme les recherches interagissent avec les objets étudiés (en quelque sorte l'idée des *interactive kind* de Hacking [Hacking, 1999]), les implications politiques et sociales de la recherche sont bien évidemment indiscutables.

Nous nous placerons ici à un niveau épistémologique, c'est-à-dire à des réflexions sur la nature et le contenu des connaissances scientifiques au sens large, c'est-à-dire co-construites et validées au sein d'une communauté imposant certains critères de scientificité [Morin, 1991], bien sûr évolutifs puisque nous nous positionnerons pour la systématisation de certains. Mais donc, même en restant à ce niveau, des prises de positions sont nécessaires, celles-ci pouvant être épistémologiques, méthodologiques, thématiques. Ces dernières ont déjà été ébauchées dans les deux chapitres précédents par les choix des objets d'étude, des problématiques, et seront renforcées à mesure de la progression.

1 https://www.cs-dc.org/

Nous proposons ainsi ici un exercice relativement original mais que nous jugeons nécessaire pour une lecture plus fluide de la suite. Il consiste en le développement précis de certains positionnements qui ont une influence particulière dans notre démarche de recherche.

Dans une première section (3.1), nous précisons notre position au regard des modèles de simulation. Après avoir détaillé les fonctions que nous prêterons aux modèles, nous argumentons sous forme d'essai pour un usage raisonné des données massives et du calcul intensif, et illustrons notre positionnement par rapport à l'exploration des modèles par une étude de cas méthodologique pour l'exploration de la sensibilité des modèles aux conditions initiales.

Dans une deuxième section (3.2), nous développons des exemples pour illustrer le besoin et la difficulté de reproductibilité, ainsi que les liens avec des nouveaux outils pouvant la favoriser mais aussi la mettre en danger. Nous illustrons la question d'ouverture des données et d'exploration interactive par une étude de cas empirique des flux de trafic en Ile-de-France.

Enfin, la dernière section (3.3) explicite modestement des positions épistémologiques, notamment concernant le courant dans lequel nous nous plaçons, la complexité des objets en sciences sociales, et la nature de la complexité de manière générale.

Le lecteur très familier avec les "commandements" de Banos [Banos, 2013] pourra trouver dans les deux premières sections des illustrations pratiques originales de ceux-ci, notre positionnement étant principalement dans leur lignée.

⋆ ⋆

⋆

*Ce chapitre est composé de divers travaux. La première section est inédite pour ses deux premières parties, et pour sa dernière partie reprend des idées présentées dans [*Cottineau *et al., 2017]. La deuxième section rend compte pour sa première partie du contenu théorique de [*Raimbault*, 2016a], et reprend [*Raimbault*, 2017b] pour l'illustration empirique. La troisième section reprend dans sa première partie les bases épistémologiques de [*Raimbault*, 2017e] approfondies par [*Raimbault*, 2017c], reprend une partie de [*Raimbault*, 2018] pour sa deuxième partie et rend compte de [*Raimbault*, 2018] pour sa dernière partie.*

3.1 MODÉLISATION, DONNÉES MASSIVES ET CALCUL INTENSIF

Nous nous positionnons à présent sur les questions liées à l'utilisation de la modélisation, des données massives et du calcul intensif, ce qui induit aussi par extension une réflexion sur les méthodes d'exploration de modèles. Il n'est pas évident que ces nouvelles possibilités soient nécessairement accompagnées de mutations épistémologiques profondes, et nous montrons au contraire que leur utilisation nécessite plus que jamais un dialogue avec la théorie. Implicitement, cette position préfigure le cadre épistémologique pour l'étude des systèmes complexes dont nous donnons le contexte en 3.3 et que nous formalisons en ouverture 8.3.

Les points développés ici couvrent certains enjeux cruciaux liés aux entreprises de modélisation, et peuvent être de nature épistémologique, théorique, ou pratique. Nous tenterons tout d'abord de répondre à la question du pourquoi de la modélisation. Nous nous positionnerons ensuite sur des questions plus techniques liées à l'utilisation des ressources de calcul émergentes et des nouvelles données. Enfin, le dernier point est méthodologique, et illustre à la fois les deux premiers tout en introduisant une nouvelle méthode d'exploration de modèles.

3.1.1 *Pourquoi modéliser ?*

Nous développons dans un premier temps le rôle de la modélisation dans notre démarche de production scientifique. Les modèles ont en apparence des rôles divers selon les disciplines : un modèle en physique découle d'une théorie, permet de la confronter à l'expérimentation et devra être validé par ses pouvoirs prédictifs avec de fortes exigences, tandis qu'en science sociales computationnelles on se contentera souvent de la reproduction de faits stylisés généraux. Un modèle statistique sera composé d'hypothèses sur des relations entre variables et sur la distribution statistique d'un terme d'erreur, et les valeurs des coefficients obtenus seront interprétées même si la mesure d'ajustement est très faible. Il s'agit donc ici de préciser dans quelle logique nos travaux de modélisation se placeront[2], quels seront leurs ressorts et objectifs.

Fonctions des modèles

Comme nous venons de l'évoquer, le terme de *modèle* a de multiples sens, et implique différentes réalités, pratiques, utilisations (on peut supposer une ontologie propre aux modèles qui deviennent des ob-

2 Si ce travail pourra paraître redondant, laborieux et superflu aux habitués des modèles de géosimulation, il est crucial dans notre logique d'ouverture disciplinaire, afin d'une part d'éviter tout malentendu sur le statut des résultats, d'autre part d'encourager un dialogue dans le cas d'utilisations très différentes des modèles.

jets réels, au moins lorsque ceux-ci sont implémentés). Une façon d'en proposer une sorte de typologie est de procéder à celle de leurs fonctions, comme le fait [Varenne, 2017], en se basant sur l'étude de diverses disciplines (biologie, géographie, sciences sociales). Cette classification est à notre connaissance la plus exhaustive qui existe. Varenne distingue donc cinq grandes classes de fonctions des modèles[3], qui vont de manière croissante dans leur intégration à une pratique sociale :

1. Fonction de perception et d'observation : rendre accessible un objet inobservable à la perception (modèle physique d'une molécule), permettre des expérimentations, une mémorisation, la lecture et la visualisation de données.

2. Fonction d'intelligibilité : description de motifs, précision des ontologies, conception par la prédiction, explication et compréhension de processus[4].

3. Fonction d'aide à la théorisation : formulation, interprétation, illustration d'une théorie, test de cohérence interne (les schémas déductifs induisent-ils des résultats de simulation de modèles contradictoires ou cohérents ?), applicabilité, calculabilité (dans le cas de schémas numériques permettant d'approcher les solutions d'équations), co-calculabilité (couplage de théories et modèles).

4. Fonction de communication sociale : communication scientifique, concertation, action avec les acteurs (*stakeholders*[5]).

5. Fonction de prise de décision : aide à la décision, action, action auto-réalisatrice dans un système abstrait (modèles de pricing en finance).

Il est clair que chaque discipline va avoir sa propre relation à ces différentes fonctions, que certaines seront privilégiées, d'autre non accessibles ou sans pertinence pour les objets étudiés ou questions posées. En physique par exemple, les aspects de validation des théories et l'existence de modèles prédictifs d'une très grande précision

3 Les grandes classes de fonctions sont déclinées en classes précises qui sont au nombre de 21. Nous ne les détaillons pas ici, mais donnons une synthèse décrivant les grandes classes.

4 La compréhension est plus générale que l'explication, car suppose une reconstruction de la structure du système et un usage déductif, c'est-à-dire une projection et génération du système considéré dans la structure psychique le considérant [Morin, 1980].

5 Nous ne développerons pas du tout cet aspect, mais tenons à préciser que les *stakeholder workshops* sont l'un des axes structurants du projet Medium que nous avons décrit en 1.3. Même si la percolation avec l'axe d'analyse et de modélisation des dynamiques des systèmes urbains dans lequel notre travail s'inscrit n'est pas explicite, celle-ci s'opèrent implicitement dans les échanges entre perspectives, et la cohabitation au sein d'un projet laisse supposer des perspectives futures plus intégrées.

sont au coeur de la discipline, tandis que des branches entières des sciences sociales comme par exemple la planification urbaine sont axées sur des modèles pour la communication et la prise de décision. A cet égard, il ne faut pas négliger la nature de science sociale de l'économie et douter des visées prédictives de certaines expériences de modélisation[6].

Cette classification des fonctions se retrouve en filigrane dans les raisons de modéliser développées en dehors de toute typologie par [Epstein, 2008] : celui-ci insiste pour tordre le cou à l'idée préconçue que les modèles servent uniquement à la prédiction, et introduit diverses raisons, parmi lesquelles on retrouve des fonctions d'intelligibilité (explication, mise en évidence de dynamiques, révéler la complexité ou la simplicité), de soutien à la théorie (découverte de nouvelles questions, mettre en valeur des incertitudes, suggérer des analogies), d'aide à la décision (solutions de crise en temps réel, trouver des compromis d'optimisation), et de communication (éduquer le public, entrainer les praticiens).

Dans ce cadre de classification fonctionnelle des modèles, notre travail utilisera principalement les fonctions suivantes :

- Modèles descriptif et extraction de motifs : il s'agira des diverses analyses empiriques visant à établir des faits stylisés sur les processus de co-évolution dans des cas d'étude donnés.

- Modèles à visée explicative et de compréhension : les modèles simulant des dynamiques territoriales que nous construirons, avec comme objectif l'intégration des processus de co-évolution, auront pour objectif principal l'explication de faits stylisés en lien avec des processus (par exemple : les variations de tel paramètre correspondant à tel processus expliquent tel fait stylisé), et dans l'idéal la *compréhension* des systèmes[7].

- Modèles pour éprouver la théorie : validation interne, c'est-à-dire cohérence du comportement du modèle par rapport aux faits stylisés impliqués par la théorie, et externe, au sens de reproduction plus ou moins performante de dynamiques de cas d'études précis considérés dans le cadre d'une théorie ; ou plus généralement pour répondre à une question ou hypothèse précise.

6 Même en finance à des fréquences élevées, où les signaux seraient plus raisonnablement assimilables à des systèmes physiques que des séries macro-économétriques par exemple comme en témoignent l'appropriation de ces problématiques par les physiciens, la prédictibilité reste questionnable et en tout cas limitée [Campbell et Thompson, 2007].

7 En fait, la frontière entre explication et compréhension est floue et subjective. Il est possible de considérer qu'il existe déjà un certain niveau de compréhension lorsqu'un modèle avec un certain niveau de cohérence interne et ontologique, en lien avec des hypothèses théoriques raisonnables et relativement autonomes, permet de tirer des conclusions sur les dynamiques globales du système considéré.

Modélisation générative

Le *type*[8] de modèles que nous utiliserons majoritairement dans notre travail s'apparente à de la *modélisation générative*, au sens donné par [Epstein, 2006] dans son manifeste pour des *sciences sociales génératives*. Le principe fondamental est de proposer d'expliquer des régularités macroscopiques comme émergentes des interactions entre entités microscopiques, en simulant l'évolution du système de manière générative[9]. Ce paradigme peut être rapproché de celui du *Pattern Oriented Modeling* en Ecologie [Grimm et al., 2005], qui vise à expliquer par production de motifs par le bas[10]. Les modèles basés-agents, c'est-à-dire des modèles impliquant un certain nombre d'agents hétérogènes relativement autonomes et simulant leur interactions, sont une façon d'y parvenir.

L'utilisation de modélisation générative peut être mise en correspondance forte avec la notion d'émergence faible introduite par [Bedau, 2002][11]. Un système qui présente des propriétés émergentes au sens faible suppose que les propriétés du niveau supérieur (macro) doivent être entièrement dérivées par simulation, tout en restant réductible sur les plans causaux et ontologiques. En d'autre termes, le niveau macro ne possède pas de pouvoirs causaux irréductibles, cela n'étant pas incompatible avec l'existence de *downward causation* et son autonomie. Certains systèmes[12] ne tombent pas dans cette

8 Dans la perspective fonctionnelle, les structures, contenus et processus, c'est-à-dire la nature des modèles en eux-mêmes (ce qui correspond à la nature et aux principes des modèles évoqués mais non classifiés par Varenne), sont donnés comme exemples en illustration, mais une fonction donnée n'est pas restreinte à un modèle donné (bien que réciproquement certains modèles ne puissent remplir certaines fonctions). Il n'existe à notre connaissance pas de typologie générale des modèles par *type*, qu'on pourrait alors définir en termes d'une typologie des relations avec les autres domaines de connaissance (voir 8.3) : par exemple un modèle utilisant telle méthodologie, privilégiant tel outil, un usage particulier ou privilégié de données, etc. Dans tous les cas, les typologies ou classification existantes de modèles sont associées aux revues de littérature et synthèses propres à chaque discipline : par exemple, [Harvey, 1969] (p. 157) propose une typologie générale qui reste toutefois inspirée de et limitée à la géographie. Les conditions de typologies interdisciplinaires sont une question ouverte, dont l'exploration dépasse largement le propos de notre travail.

9 En gardant à l'esprit que la capacité à générer est bien sûr une composante nécessaire mais pas suffisante à l'explication, comme l'illustre le débat à ce sujet autour des travaux d'Epstein synthétisé par [Rey-Coyrehourcq, 2015] (p. 154).

10 En effet, le POM vise à ce que le modèle reproduise par la simulation, c'est-à-dire *génère*, des motifs (*patterns*) attendus à différentes échelles, constituant un laboratoire virtuel dans lequel des hypothèses peuvent être testées. Par ailleurs, la générativité d'Epstein se base sur des paradigmes similaires pour l'explication, impliquant des modèles à la complexité progressive et qui permettent le test d'hypothèses, en isolant des mécanismes suffisant pour reproduire des motifs macroscopiques.

11 On rappelle que l'émergence faible correspond à l'émergence de propriété à un niveau supérieur qui doivent être effectivement computées par le système pour être connues.

12 Comme le montrent la conscience en neuroscience et psychologie, ou les débats sur l'existence et l'autonomie "d'êtres sociétaux" en sociologie [Angeletti et Berlan,

catégorie à notre connaissance actuelle puisqu'on n'est pas capable de désigner des éléments microscopiques causaux. En revanche, des systèmes qu'on comprend mal mais qui se simulent eux-mêmes et dont on est certain que l'état macro émerge des interactions microscopiques (prenons le traffic et la congestion par exemple), sont des parfaites illustrations de cette notion. Les exemples donnés par Bedau pour démontrer son propos sont des automates cellulaires en deux dimensions, pour lesquels le rôle de la computation est évident et la *downward causation* peut être illustrée par le comportement des structures macroscopiques du jeu de la vie qui agissent rétroactivement sur les cellules. Connaitre les dynamiques de systèmes faiblement émergents nécessite par définition de les simuler, et donc de les modéliser[13], cette approche est ainsi naturelle pour connaître la structure ou les processus dans un système complexe.

Le modèle comme outil de connaissance indirecte

Ainsi, nos modèles seront principalement à visée de comprehension (même s'ils n'atteignent pas l'objectif et restent au niveau d'une explication). Nous procéderons dans certains cas à des calibrations fines sur données observées, mais celles-ci n'auront à aucun moment l'objectif de prédiction. Ces calibrations serviront à extrapoler des paramètres et apprendre indirectement sur les processus modélisés, et le modèle est ainsi bien un instrument de *connaissance indirecte*.

Cette connaissance des processus est permise par l'utilisation de la simulation comme un laboratoire virtuel permettant le test d'hypothèses formulées à partir d'une théorie ou issues de faits stylisés empiriques : c'est exactement ce type de paradigme que construit [Pumain et Reuillon, 2017e], qui insiste sur (i) le besoin de parcimonie dans les modèles ; (ii) le besoin de multiples modèles (multi-modélisation) ; et (iii) le rôle de l'exploration extensive des modèles, pour y parvenir sans tomber dans le piège de l'équifinalité[14]. Ainsi, l'établissement

2015], pour lesquels nous pourrions ne connaitre que partie seulement des éléments causaux microscopiques à savoir les individus.

13 Sur la différence entre simulation de modèle et modèle de simulation, [Phan et Varenne, 2010] explique dans quelle mesure ces deux notions peuvent être distinguées, mais que cela n'implique pas de différence fondamentale pour l'application concrète : la simulation d'un modèle consiste en l'opération de computation des états successifs d'un modèle dans une configuration donnée, tandis qu'un modèle de simulation est un modèle conçu pour la simulation d'un système (par exemple un modèle génératif) ou la simulation d'un autre modèle (par exemple les schémas numériques pour approcher des équations). Dans tous les cas, l'utilisation du modèle impliquera simulation d'un modèle. Ces remarques se vérifient particulièrement dans le cas de la modélisation générative. Nous utiliserons les deux de manière interchangeable par la suite.

14 L'équifinalité correspond à la possibilité pour un système d'atteindre un point de son espace des phases par des trajectoires différentes, c'est-à-dire dans notre cas des motifs macroscopiques pouvant être générés par différents processus microscopiques. Ce concept était déjà formulé dans la théorie générale des systèmes [Von Bertalanffy, 1972]. Il pose problème aux notions de causalité, et remet en cause

des *Calibration Profiles* du modèle SimpopLocal [Reuillon et al., 2015] permettent d'établir des conditions nécessaires et suffisantes pour reproduire un motif donné, et donc par exemple de déclarer indirectement un processus nécessaire ou non pour produire un fait stylisé.

Ainsi, nous prendrons ici ce parti de l'utilisation des modèles (de simulation principalement), tout en gardant à l'esprit que celui-ci ne répond que partiellement aux challenges fondamentaux de la modélisation urbaine donnés par [Perez, Banos et Pettit, 2016], notamment la capture de la complexité et de la multidimensionalité des systèmes urbains ainsi que la possibilité de générer des scenarii futurs possibles (ce qui est différent de la prédiction), mais pas la question des modèles de planification urbaine, pouvant par exemple être participatifs et impliquant les *stakeholders*[15].

Comment explorer un modèle de simulation

Afin d'éviter au maximum le "bricolage" concernant l'ensemble des étapes de la genèse d'un modèle, de sa spécification, sa conception, son utilisation à son exploration, décrit par [Kotelnikova-Weiler et Le Néchet, 2017], nous proposons de nous fixer un protocole pour la partie d'exploration des modèles. Plus généralement, il existe des protocoles généraux comme celui introduit par [Grimm et al., 2014] pour accompagner l'ensemble de la démarche de modélisation. Nous considérons l'étape d'exploration et creusons celle-ci plus en détails. Nous nous plaçons dans le cadre fixé ci-dessus d'un modèle de simulation, majoritairement à visée de compréhension.

Le protocole simplifié est issu directement de la philosophie et de la structure d'OpenMole. On peut se référer par exemple à [Reuillon, Leclaire et Rey-Coyrehourcq, 2013] pour les principes fondamentaux, la documentation en ligne[16] pour un aperçu global des méthodes disponibles et de leur articulation dans un cadre standard, et [Pumain et Reuillon, 2017b] pour une contextualisation des différentes méthodes. Ces travaux[17] ont apporté un nombre considérable d'innovations à la fois méthodologiques, techniques, thématiques et théoriques. La philosophie d'OpenMole s'articule autour de trois axes (voir entretien avec R. Reuillon, Annexe D.3) : le modèle comme "boite noire" à explorer (i.e. méthodes indépendantes du modèle), utilisation de méthodes avancées d'exploration, accès transparent aux environnements de calcul intensif. Ces différentes composantes sont

des explications de causalité "directe" au niveau macroscopique - nous y reviendrons plus particulièrement en 4.2.

15 Le rôle de la visée d'application des modèles est lié à la fois à une sensibilité disciplinaire, comme le domaine des Luti [Wegener et Fürst, 2004] qui l'est bien plus que celui de la géographie théorique et quantitative, mais aussi à une sensibilité "culturelle", comme l'illustre [Batty, 2013b] qui montre une branche de la géographie anglo-saxonne plus proche des applications concrètes.

16 Disponible à `https://next.openmole.org/Models.html`.

17 La majorité ayant été réalisés dans le cadre interdisciplinaire de l'ERC Geodivercity.

en interdépendance forte, et permettent un changement de paradigme dans l'utilisation des modèles de simulation : utilisation de multi-modélisation, c'est-à-dire structure variable du modèle [Cottineau et al., 2015], changement de la nature des questions posées au modèle (par exemple détermination complète de l'espace faisable [Chérel, Cottineau et Reuillon, 2015]), tout cela permis par l'utilisation du calcul intensif [Schmitt et al., 2015].

Nous considérons un modèle de simulation comme un algorithme produisant des sorties à partir de données et de paramètres en entrée. Dans ce cadre, nous proposons dans un cas idéal l'ensemble des étapes suivantes qui devraient être nécessaire pour une utilisation robuste des modèles de simulation.

1. Identification des mécanismes principaux et des paramètres cruciaux associés, possiblement des méta-paramètres (ici compris comme paramètre générant la configuration initiale du modèle), ansi que de leur domaine thématique ; identification des indicateurs pour évaluer la performance ou le comportement du modèle.

2. Évaluation des variations stochastiques : grand nombre de répétitions pour un nombre raisonnable de paramètres, établissement du nombre de répétitions nécessaire pour atteindre un certain niveau de convergence statistique.

3. Évaluation de la sensibilité aux meta-paramètres, suivant la méthodologie innovante développée par la suite[18].

4. Exploration brutale pour une première analyse de sensibilité, si possible evaluation statistique des relations entre paramètres et indicateurs de sortie.

5. Calibration, exploration algorithmique ciblée par l'utilisation d'algorithmes spécifiques (*Calibration Profile*, *Pattern Space Exploration*)[19]

6. Retours sur le modèle, extension et nouvelles briques de multi-modélisation, retours sur les faits stylisés et la théorie.

Le cas échéant, certaines étapes n'ont pas lieu d'être, par exemple l'évaluation de la stochasticité dans le cas d'un modèle déterministe. De même, les étapes prendront plus ou moins d'importance selon la nature de la question posée : la calibration ne sera pas pertinente dans le cas de modèles complètement synthétiques, tandis qu'une

18 Un exemple de cette méthodologie consistant à la génération de données synthétiques sera utilisé dans la suite de cette section ; une description formelle de la méthode est donnée en B.3 et un autre exemple d'application en C.3.

19 Nous ne pratiquerons quasiment pas ce dernier point, trouvant suffisamment de réponses à nos questions avec les points précédents.

exploration systématique d'un grand nombre de paramètres ne sera pas forcément nécessaire dans le cas d'un modèle qui a pour but d'être calibré sur des données.

Lien entre modélisation et science ouverte

Enfin, il est important de souligner brièvement les liens entre pratiques de modélisation et science ouverte, en parallèle du lien entre reproductibilité et science ouverte que nous ferons à la fin de 3.2. En fait, la science ouverte est composée d'un ensemble de pratiques se déclinant sur différents domaines, d'où sa ventilation logique dans nos positionnements. Pour illustrer les enjeux, nous proposons de décrire l'exemple des workflows d'exploration de modèle comme une méthode de méta-analyse de sensibilité, c'est-à-dire un aspect de la méthodologie appliquée ci-dessus.

Les idées de multi-modélisation et d'exploration intensive de modèle sont tout sauf nouvelles puisque Openshaw défendait déjà le "model-crunching" dans [Openshaw, 1983], mais leur utilisation effective commence seulement à émerger grâce à l'apparition de nouvelles méthodes et outils en même temps qu'une explosion des capacités de calcul : [Cottineau, Rey et Reuillon, 2016] propose une approche renouvelée de la multi-modélisation. Le couplage de modèles tel que nous l'opérons répond à des questions similaires. Dans cette lignée de recherche, la plateforme d'exploration de modèle OpenMole [Reuillon, Leclaire et Rey-Coyrehourcq, 2013] permet d'embarquer n'importe quel modèle comme une boîte noire, d'écrire des workflow d'exploration modulables qui utilisent des méthodologies d'exploration avancées comme des algorithmes génétiques, et de distribuer de manière transparente les calculs sur des infrastructures de calcul à grande échelle comme des clusters ou grilles de calcul. Dans le cas précédent, l'outil du workflow est un outil puissant pour intégrer à la fois l'analyse de sensibilité et la méta-analyse de sensibilité, et permet de coupler n'importe quel générateur avec n'importe quel modèle de façon très directe, à la condition minimale que le modèle puisse être paramétré sur sa configuration spatiale initiale, par la donnée de méta-paramètres ou d'une configuration entière.

D'autre part, une idée des workflow est de favoriser des constructions ouvertes et collaboratives, puisque le "marketplace" d'OpenMole, directement intégré au logiciel[20], permet de bénéficier directement des exemples qui auront été partagés sur le dépôt collaboratif. Cela ressemble aux plateformes de partage de modèles, qui sont nombreuses pour les modèles agents par exemple, mais dans un esprit encore plus modulaire et participatif. Ainsi, certains choix épistémologiques et méthodologiques au regard de la modélisation impliquent directement un positionnement au regard de la science ouverte : la

20 autrement accessible à `https://github.com/openmole/openmole-market`

multi-modélisation et les familles de modèles, qui vont de pair avec le couplage de modèle hétérogènes et multi-échelles, ne peuvent guère être viables sans des pratiques d'ouverture, de partage et de construction collaborative des modèle, comme le rappelle [Banos, 2013].

Enfin, l'un des visages de la construction de connaissances ouvertes est la pédagogie. [Chen et Levinson, 2006] propose la simulation comme outil pour apprendre aux élèves ingénieurs les processus sous-jacents aux systèmes qu'ils seront amenés à concevoir et gérer. Cet aspect est également à garder en tête de par son caractère performatif : les modèles ont alors une retroaction sur les situations réelles, ce qui complexifie encore le système considéré.

Synthèse

Résumons brièvement les idées à garder à l'esprit à la suite de ce survol rapide d'enjeux cruciaux liés à la modélisation.

1. Les modèles peuvent avoir un grand nombre de fonctions [Varenne, 2017], parmi lesquelles nous utiliserons fondamentalement : extraction d'information et de motifs, explication et compréhension, vérification et construction des théories.

2. Nous nous placerons majoritairement dans le paradigme de la *modélisation générative*, dans un souci de parcimonie et de modèles multiples avec des protocoles d'exploration extensive appropriés [Pumain et Reuillon, 2017e].

3. Cette façon de modéliser à la fois suppose et participe à une démarche de science ouverte [Fecher et Friesike, 2014].

Dans ce contexte, nous proposons de développer à présent certains enjeux particulièrement important pour notre question de manière plus précise.

3.1.2 *Pour un usage raisonné des données massives et de la computation*

La *révolution des données massives* réside autant dans la disponibilité de grands jeux de données de nouveaux types variés, que dans la puissance de calcul potentielle toujours en augmentation. Même si le *tournant computationnel* ([Arthur, 2015]) est central pour une science consciente de la complexité et est sans douter la base des pratiques de modélisation futures en géographie comme [Banos, 2013] souligne, nous soutenons que à la fois le *déluge de données* et les *capacités de calcul* sont dangereuses si non cadrées dans un cadre théorique et formel propre. Le premier peut biaiser les directions de recherche vers les jeux de données disponibles avec le risque de se déconnecter d'un fond théorique, tandis que le second peut occulter des résolutions analytiques préliminaires essentielles pour un usage cohérent

des simulations. Nous avançons que les conditions pour la majorité des résultats dans cette thèse sont en effet ceux mis en danger par un enthousiasme inconsidéré pour les données massives, tirant la conclusion qu'un challenge majeur pour la géocomputation future est une intégration sage des nouvelles pratiques au sein du corpus existant de connaissances.

Accroissement de la puissance de calcul

La puissance de calcul disponible semble suivre un tendance exponentielle, comme une sorte de loi de Moore. Grace à d'une part la loi de Moore effective pour le matériel, d'autre part l'amélioration des logiciels et algorithmes, conjointement avec une démocratisation de l'accès au infrastructures de simulation à grande échelle, permet à toujours plus de temps processeur d'être disponible pour le chercheur en sciences sociales (et pour le scientifique en général, mais cette mutation a déjà été opérée depuis plus longtemps dans d'autres domaines). Il y a environ une dizaine d'années, [Gleyze, 2005] était forcé de conclure que les analyses de réseau, pour les transports publics parisiens, étaient "limitées par le calcul". Aujourd'hui la plupart des mêmes analyses seraient rapidement réglée sur un ordinateur personnel avec les logiciels et programmes appropriés : [Lagesse, 2015] est un témoin d'un tel progrès, introduisant des nouveaux indicateurs avec une plus grande complexité de calcul, qui sont calculés sur des réseaux à grande échelle. Le même parallèle peut être fait pour les modèles Simpop : les premiers modèles Simpop au début du millénaire [Sanders et al., 1997] étaient "calibrés" à la main, tandis que [Cottineau et al., 2015] calibre le modèle Marius en multi-modélisation et [Schmitt et al., 2015] calibre très précisément le modèle SimpopLocal, chacun sur la grille avec des milliards de simulations. Un dernier exemple, le champ de la *Space Syntax*, a témoigné d'une longue route et de progrès considérables depuis ses origines théoriques [Hillier et Hanson, 1989] jusqu'à ses récentes applications à grande échelle [Hillier, 2016].

Un déluge de données ?

Concernant les nouvelles données "massives" qui sont disponibles, il est clair que des quantités toujours plus grandes et des types toujours nouveaux sont disponibles. De nombreux exemples de champs d'application peuvent être donnés. La mobilité en est typique, puisque étudiée selon divers points de vue, comme les nouvelles données issues des systèmes de transport intelligents [O'brien, Cheshire et Batty, 2014], des réseaux sociaux [Frank et al., 2014], ou des données plus exotiques comme des données de téléphonie mobile [De Nadai et al., 2016]. Dans un autre esprit, l'ouverture de jeux de données "classiques" (comme les applications synthétiques urbaines, les initiatives

gouvernementales pour les données ouvertes) devrait permettre toujours plus de méta-analyses. De nouvelles façons de pratiquer la recherche et produire des données sont également en train d'émerger, vers des initiatives plus interactives et venant de l'utilisateur. Ainsi, [COTTINEAU, 2017] décrit une application web ayant pour but de présenter une méta-analyse de la loi de Zipf sur de nombreux jeux de données, mais en particulier inclut une option de dépôt, à travers laquelle l'utilisateur peut télécharger son propre jeu de données et l'inclure dans la méta-analyse. D'autres applications permettent l'exploration interactive de la littérature scientifique pour une meilleure connaissance d'un horizon scientifique complexe, comme [CHASSET et al., 2016] fait.

Des dangers induits

Comme toujours la situation n'est naturellement pas aussi idyllique qu'elle semble être au premier abord, et l'herbe verte du pré du voisin que nous pouvons être tentés d'aller brouter se transforme rapidement en un triste fumier. En effet, les objectifs et motivations d'un grand nombre d'approches restent flous et on peut facilement s'y perdre. Des illustrations parleront d'elles-mêmes.

[BARTHELEMY et al., 2013] introduit un nouveau jeu de données et des méthodes relativement nouvelles pour quantifier l'évolution du réseau de rues, mais les résultats, sur lesquels les auteurs semblent s'étonner, sont qu'une transition a eu lieu à Paris à l'époque d'Haussmann. Tout historien de l'urbanisme s'interrogerait sur le but exact de l'étude, puisqu'il reste à la fin un sentiment de réinvention de la roue. L'utilisation des ressources de calcul peut également être exagéré, et dans le cas de la modélisation multi-agents, on peut citer [AXTELL, 2016], pour lequel l'objectif de simuler le système à l'échelle 1 :1 semble être loin des motivations et justifications originelles de la modélisation multi-agents, et pourrait même donner des arguments aux économistes *mainstream* qui dénigrent facilement les ABMS.

D'autres anecdotes peuvent inquiéter : il existe en ligne des exemples étonnants, comme une application web[21] qui utilise des ressources de calcul pour simuler des distributions Gaussiennes afin de calculer pour un modèle de Gibrat les moyenne et variance, qui sont des paramètres d'entrée du modèle. En résumé, cela revient à vérifier le Théorème de la Limite Centrale. D'autre part, la distribution complète donnée par un modèle de Gibrat est entièrement connue théoriquement comme résolu e.g. par [GABAIX, 1999].

Sur ce point, nous devons partiellement être en désaccord avec le neuvième commandement de BANOS, qui rappelle que "les mathématiques ne sont pas le language universel des modèles", ou plutôt sou-

21 Voir `http://shiny.parisgeo.cnrs.fr/gibratsim/`.

ligner les dangers d'une mauvaise interprétation de ce principe[22] : il postule que des moyens alternatifs aux mathématiques existent pour faire comprendre des processus ou des méthodes, mais précise que ceux-ci sont une porte d'entrée et ne prétend jamais qu'il est possible de se passer des mathématiques, dérive que l'exemple précédent illustre parfaitement. D'ailleurs, il est possible d'exhiber des structures mathématiques très simples, comme un simplexe en dimension quelconque, dont la visualisation "simple" est un problème ouvert.

Les données fournissent aussi leur collection de dérives. Récemment, sur la liste de diffusion de géographie francophone *Geotamtam*, un soudain engouement autour des données issues de *Pokemon Go* a semblé répondre plus à un besoin urgent et inexpliqué d'exploiter cette source de données avant tous les autres, plutôt qu'à des considérations théoriques élaborées. Des jeux de données existant et précis, comme la population historiques des villes (pour la France la base Pumain-INED par exemple), sont loin d'être entièrement exploités et il pourrait être plus pertinent de se concentrer sur ces jeux de données classiques qui existent déjà. De même, il faut être conscient des possibles applications de résultats basée sur des malentendus : [Louail et al., 2017] analyse la redistribution potentielle des transactions de carte bancaire au sein d'une ville, mais présente les résultats comme la base possible de recommandations de politiques pour une équité sociale en agissant sur la mobilité, oubliant que la forme et les fonctions urbaines sont fortement couplées et que déplacer des transactions d'un endroit à un autre implique des processus bien plus complexes que des régulations directes, qui d'autant plus ne s'appliquent jamais de la façon prévue et conduisent à des résultats différents de ceux attendus. Une telle attitude, souvent observée de la part de physiciens, est très bien mise en allégorie par la figure 12 qui n'est qu'à moitié une exagération de certaines situations.

Pour un usage raisonné

Notre principal argument est que le tournant computationnel et les pratiques de simulation seront centrales en géographie, mais peuvent également être dangereux, pour les raisons illustrées ci-dessus, i.e. que le déluge de données peut imposer les sujets de recherche et occulter la théorie, et que la computation peut éluder la construction et la résolution de modèles. Un lien plus fort est nécessaire entre les pratiques de calcul, l'informatique, les mathématiques, les statistiques et la géographie théorique.

La géographie théorique et quantitative est au centre de cette dynamique, puisqu'il s'agit de sa motivation initiale principale qui semble

22 De manière générale, les commandements de Banos paraissent simples dans leur formulation, mais sont d'une profondeur et d'une complexité déconcertante lorsqu'on essaye d'en tirer les implications et la philosophie globale sous-jacente, et ne doivent jamais être pris à la légère.

oubliée dans certains cas. Cela implique un besoin de recherche de théories élaborées intégrées avec des pratiques de simulation conscientes. En d'autres mots, on peut répondre à des questions naïves complémentaires qui ont toutefois besoin d'être traitées une bonne fois pour toutes. Quant à la question de la possibilité d'une géographie quantitative libérée de la théorie, la réponse est naturellement négative puisque cela se rapproche du piège de la fouille de données par boîte noire. Quoi qu'il soit fait par cette approche, les résultats auront un pouvoir explicatif très faible, puisqu'ils pourront mettre en valeur des relations mais pas reconstruire des processus. D'autre part, la possibilité d'une géographie quantitative purement basée sur le calcul est une vision dangereuse : même le gain de trois ordres de grandeur dans la puissance de calcul disponible ne résout pas le sort de la dimension.

Prenons l'exemple des résultats de non-stationnarité obtenus en 4.1. L'utilisation de données relativement massives, de par les algorithmes spécialement conçus pour être capable de faire les traitements, est une condition nécessaire aux résultats obtenus, mais à la fois l'échelle et les objets (c'est-à-dire les indicateurs calculés) sont co-déterminés par les constructions théoriques. En effet l'absence de théorie impliquerait de ne pas connaitre les objets, mesures et propriétés à étudier (e.g. le caractère multi-scalaire ou dynamique des processus), et sans résolutions analytiques, il serait souvent difficile de tirer des conclusions à partir des analyses empiriques seules, notamment pour l'aspect multi-scalaire.

Rien n'est vraiment nouveau ici mais cette position doit être affirmée et tenue, précisément car notre travail se base sur ce type d'outils, essayant d'avancer sur une arête fine et fragile, avec d'un côté le vide du charlatanisme théorique infondé et de l'autre l'abîme de l'overdose technocratique dans des quantités de données folles. Plus que jamais on a besoin de théories simples mais fondées et puissantes à-la-Occam [Batty, 2016], pour permettre une intégration saine des nouvelles techniques au sein des connaissances existantes.

3.1.3 *Étendre les analyses de sensibilité*

Contexte

Lors de l'évaluation de modèles basés sur les données, ou même de modèles plus simples partiellement basés sur les données impliquant une paramétrisation simplifiée, une issue inévitable est le manque de contrôle sur les "paramètres implicites du systèmes" (ce qui n'est pas une notion stricte mais doit être compris dans notre sens comme les paramètres régissant la dynamique). En effet, une statistique issue d'executions du modèle sur un nombre suffisant d'executions peut toutefois rester biaisée, au sens où il est impossible de savoir si les résultats sont dus aux processus que le modèle cherche à traduire

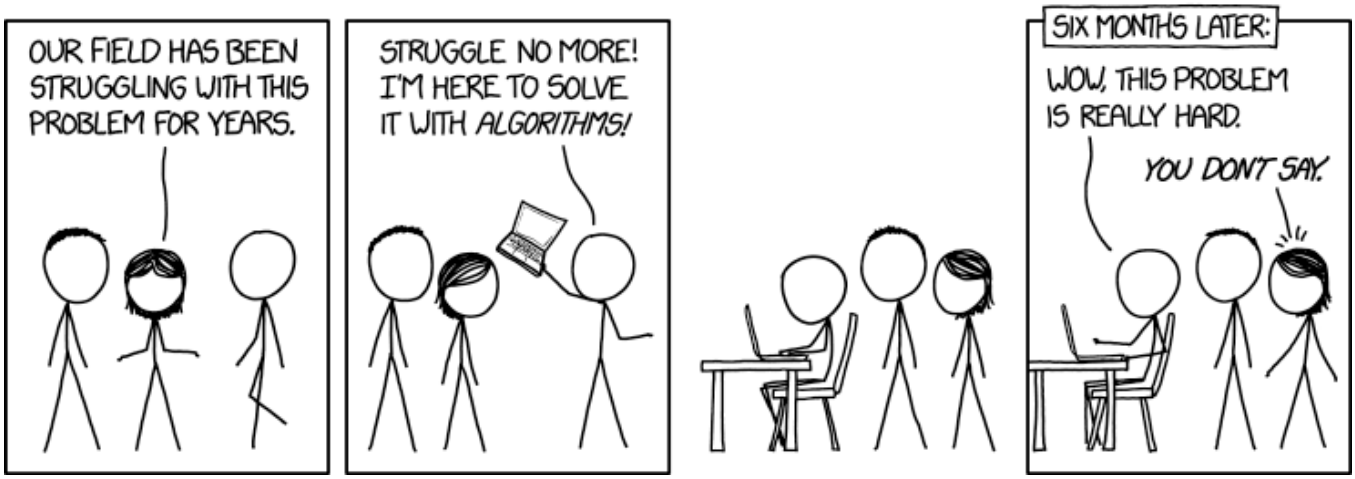

FIGURE 12 : De l'usage naïf de la fouille de données et du calcul intensif. Source : xkcd

ou à une structure présente dans les données initiales. La question méthodologique fondamentale qui nous intéressera pour la suite est d'être capable d'isoler les effets propres aux processus du modèles de ceux liés à la géographie.

CONTEXTE Bien que les modèles de simulation des systèmes géographiques en général et les modèles basés-agent en particulier représentent une opportunité considérable d'explorer les comportements socio-spatiaux et de tester une variété de scenarios pour les politiques publiques, la validité des modèles génératifs est incertaine tant que la robustesse des résultats n'a pas été établie. Les analyses de sensibilité incluent généralement l'analyse des effets de la stochasticité sur la variabilité des résultats, ainsi que les effets de variations locales des paramètres. Cependant, les conditions spatiales initiales sont généralement prise pour données dans les modèles géographiques, laissant ainsi totalement inexploré l'effet des motifs spatiaux sur les interactions des agents et sur leur interaction avec l'environnement. Dans cette partie, nous présentons une méthode pour établir l'effet des conditions spatiales initiales sur les modèles de simulation, utilisant un générateur systématique contrôlé par des meta-paramètres pour créer des grilles de densité utilisées dans les modèles de simulation spatiaux. Nous montrons, avec l'exemple d'un modèle agent très classique (le modèle Sugarscape d'extraction de ressources) que l'effet de l'espace dans les simulations est significatif, et parfois plus grand que l'effet des paramètres eux-mêmes. Nous y arrivons en utilisant le calcul haute performance en un workflow très simple et open source. Les bénéfices de notre approche sont variés mais incluent par exemple la connaissance du comportement du modèle dans un contexte plus large, la possibilité de contrôle statistique pour régresser les sorties

du modèles, ou une exploration plus fine des dérivées du modèle que par rapport à une approche directe.

ROLE DE LA DÉPENDANCE AU CHEMIN SPATIO-TEMPORELLE La dépendance au chemin spatio-temporelle est une des raisons principales rendant notre approche pertinente. En effet, un aspect crucial de la plupart des systèmes complexes spatio-temporels est leur non-ergodicité [PUMAIN, 2012b] (la propriété que les échantillons dans l'espace ne sont pas équivalent aux échantillons dans le temps pour calculer des statistiques comme la moyenne), qui témoigne généralement de forte dépendances au chemin spatio-temporelles dans les trajectoires. De manière similaire à ce que GELL-MANN appelle *frozen accidents* dans tout système complexe [GELL-MANN, 1995], une configuration donnée contient des indices sur les bifurcations passées, qui peuvent avoir eu des effets considérables sur l'état du système. Les effets temporels et cumulatifs ont été considérés dans de nombreux sous-champs géographiques et à différentes échelles géographiques, par exemple les systèmes régionaux [WILSON, 1981] ou l'échelle intra-urbaine [ALLEN et SANGLIER, 1979]. L'impact de la configuration spatiale sur les dynamiques du modèle et les bifurcations spatiales a été moins étudié.

L'exemple des réseaux de transport est une bonne illustration, car leur forme spatiale et leur hiérarchie est fortement influencée par les décisions d'investissement du passé, les choix techniques, ou des décisions politiques qui ne sont parfois pas rationnelles [ZEMBRI, 2010]. Certains indicateurs agrégés ne prendront pas en compte les positions et trajectoires de chaque agent (comme les inégalités totales dans le modèle Sugarscape) mais d'autres, comme dans le cas des motifs d'accessibilité spatiale dans un système de villes, capture entièrement la dépendance au chemin et peuvent ainsi être fortement dépendants à la configuration spatiale initiale. Il n'est pas clair par exemple ce qui a causé la transition de la capitale française de Lyon à Paris dans le bas Moyen-Age, certaines hypothèses étant la reconfiguration des motifs commerciaux du Sud au Nord de l'Europe et donc une centralité accrue pour Paris due à sa position spatiale, tout en gardant à l'esprit que les centralités géographique et politique ne sont pas équivalentes et entretiennent une relation complexe [GUENÉE, 1968]. La bifurcation induite par des facteurs socio-économiques et politiques a pris une signification profonde avec des répercussions mondiales encore aujourd'hui quand elle a été concrétisée par la configuration spatiale.

TRAVAUX EXISTANTS L'effet de la configuration spatiale sur les attributs agrégés à la zone des comportements humains a été largement discuté en géostatistiques, approximativement depuis l'introduction du *Modifiable Areal Unit Problem* (MAUP) [OPENSHAW, 1984]. Plus ré-

cemment, [Kwan, 2012] plaide pour un examen plus attentif de ce qui serait un *Uncertain Geographic Context Problem* (UGCoP), qui est la configuration spatiale des unités géographiques même si la taille et la délimitation des zones est la même. Au contraire, le faible nombre de considérations similaires dans la littérature traitant des modèles de simulation géographiques remet en question la généralisation de leur résultats, comme cela a été montré par exemple dans le cas des modèles LUTI [Thomas et al., 2018], ou des processus de diffusion étudiés par modèles multi-agents [Le Texier et Caruso, 2017].

Méthodes

Nous détaillons à présent la méthode développée pour analyser la sensibilité des modèles de simulation aux conditions spatiales initiales. S'ajoutant au protocole usuel, qui consiste à simuler un modèle μ pour différentes valeurs de ses paramètres et faire le lien entre ces variations aux variations des résultats de simulation, nous introduisons ici un générateur spatial, qui est lui-même déterminé par des paramètres et produit des ensembles de configurations spatiales initiales. Les configurations spatiales initiales sont catégorisées pour représenter des types d'espace typiques (par exemple des grilles de densité monocentriques ou polycentriques), et la sensibilité du modèle est à présent testée sur les paramètres de μ mais aussi sur les paramètres spatiaux ou les types spatiaux. Cela permet à l'analyse de sensibilité de fournir des conclusions qualitatives au regard de l'influence de la distribution spatiale sur les sorties des modèles de simulation, en parallèle des variation classiques des paramètres.

GÉNÉRATEUR SPATIAL Le générateur spatial applique un modèle de morphogenèse urbaine développé et exploré en 5.2. Pour le présenter rapidement, les grilles sont générées par un processus itératif qui ajoute une quantité de population N_G à chaque pas de temps, l'allouant selon un attachement préférentiel caractérisé par sa force d'attraction α. Le premier processus est ensuite lissé n_d fois par un processus de diffusion de force β. Les grilles sont donc générées aléatoirement par la combinaison des valeurs de ces quatre meta-paramètres α, β, n_d and N_G. Pour faciliter l'exploration, seule la distribution de densité est autorisée à varier plutôt que la taille de la grille, qui est fixée à un environnement carré 50x50 de population 100,000 unités.

COMPARER LES DIAGRAMMES DE PHASE Afin de tester l'influence des conditions spatiales initiales, nous avons besoin d'une méthode systématique pour comparer des diagrammes de phase. En effet, nous avons autant de diagramme de phase que de grilles spatiales, ce qui rend une comparaison visuelle qualitative non réaliste. Une solution est d'utiliser des procédures quantitatives systématiques. De nombreuses méthodes pourraient potentiellement être utilisées : par

exemple, des indicators anisotropes comme la donnée de clusters et leur position dans le diagramme de phase, peuvent permettre de révéler des *meta-transitions de phase* (transition de phase dans l'espace des meta-paramètres). L'utilisation de métriques comparant des distributions spatiales, comme la *Earth Movers Distance* qui est utilisée en vision par ordinateur pour comparer des distributions de probabilité [Rubner, Tomasi et Guibas, 2000], ou la comparaison de matrices de transition agrégées de la dynamique associée au potentiel décrit par chaque distribution, est également possible. Les méthodes de comparaison de cartes, répandues en sciences environnementales, fournissent de nombreux outils pour comparer des champs en deux dimensions [Visser et De Nijs, 2006]. Pour comparer un champ spatial évoluant dans le temps, des méthodes élaborées comme les Fonctions Orthogonales Empiriques qui isolent les variations temporelles des variations spatiales, seraient applicables dans notre cas en prenant le temps comme une dimension de paramètre, mais celles-ci ont été montrées ayant une performance similaire à la comparaison visuelle directe lorsqu'on prend la moyenne sur un ensemble de contributions crowdsourcées [Koch et Stisen, 2017]. Pour rester simple et car de telles considérations méthodologiques sont auxiliaires pour le propos principal de cette partie, nous proposons une mesure intuitive correspondant à la part de la variabilité inter-diagrammes relativement à leur variabilité interne. Plus formellement, cette distance est donnée par

$$d_r(\alpha_1, \alpha_2) = 2 \cdot \frac{d(f_{\vec{\alpha_1}}, f_{\vec{\alpha_2}})^2}{Var\left[f_{\vec{\alpha_1}}\right] + Var\left[f_{\vec{\alpha_2}}\right]} \tag{1}$$

où $\alpha \mapsto [\vec{x} \mapsto f_{\vec{\alpha}}(\vec{x})]$ est l'opérateur donnant les diagrammes de phase avec $\vec{x}$ paramètres et $\vec{\alpha}$ meta-paramètres, et d une distance entre distributions de probabilité qui peut être prise par exemple comme la distance L2 basique ou la *Earth Movers Distance*. Pour chaque valeur $\vec{\alpha_i}$, le diagramme de phase est vu comme un champ spatial aléatoire, ce qui facilite la définition des variances et de la distance.

Résultats

Sugarscape est un modèle d'extraction de ressources qui simule la distribution inégale des richesses dans une population hétérogène [Epstein et Axtell, 1996]. Des agents ayant différentes portées de vision et différents métabolismes collectent une ressource qui se régénère automatiquement et disponible de manière hétérogène dans le paysage initial. Ceux-ci s'établissent et collectent la ressource, ce qui mène certains d'entre eux à survivre et d'autres à périr. Les paramètres principaux du modèle sont le nombre d'agents, leur ressources minimale et maximale. Nous nous intéressons principalement à tester l'impact de la distribution spatiale, en utilisant le générateur spatial.

La sortie du modèle est mesurée comme le diagramme de phase d'un index d'inégalité pour la distribution de la ressource (index de Gini). Nous étendons l'implémentation ayant initialement une distribution de richesse des agents, donnée par [Li et Wilensky, 2009].

Pour l'exploration, $2.5 \cdot 10^6$ simulations (1000 points de paramètres x 50 grilles de densité x 50 réplications) nous permettent de montrer que le modèle est bien plus sensible à l'espace qu'à ses autres paramètres, à la fois quantitativement et qualitativement : l'amplitude des variations entre les grilles de densité est plus grande que l'amplitude dans chaque diagramme de phase, et le comportement de ces diagrammes de phase est qualitativement différent dans diverses régions de l'espace morphologique. Plus précisément, nous explorons une grille d'un espace de paramètre basique du modèle, dont les trois dimensions sont la population des agents $P \in [10; 510]$, la ressource minimale initiale par agent $s_- \in [10; 100]$ et la ressource initiale maximale par agent $s_+ \in [110; 200]$. Chaque paramètre est discrétisé en 10 valeurs, donnant 1000 points de paramètres. Nous procédons à 50 répétitions pour chaque configuration, ce qui donne des propriétés de convergence raisonnables. La distribution spatiale initiale varie parmi 50 grilles initiales, générée en échantillonnant les méta-paramètres du générateur dans un Hypercube Latin. Nous démontrons ainsi la flexibilité de notre cadre, par le couplage séquentiel direct du générateur avec le modèle. Nous mesurons la distance de l'ensemble des diagrammes de phase à 3 dimensions à un diagramme de phase de référence calculé sur l'initialisation du modèle par défaut (voir Fig. 13 pour sa position morphologique au regard des grilles générées), en utilisant l'équation 1 avec la distance L2 pour assurer une interprétation directe. En effet, cela donne dans ce cas la distance au carré moyenne entre chaque point en correspondance des diagrammes, relative à la moyenne des variances de chaque. Pour cela, des valeurs plus grandes que 1 signifient que la variabilité inter-diagramme est plus importante que la variabilité intra-diagramme.

Nous obtenons une sensibilité très forte aux conditions initiales, puisque la distribution de la distance relative à la référence s'étend sur l'ensemble des grilles de 0.09 à 2.98, avec un médiane de 1.52 et une moyenne de 1.30. Cela signifie qu'en moyenne, le modèle est plus sensible aux méta-paramètres qu'aux paramètres, et que la variation relative peut atteindre jusqu'à un facteur 3. Nous montrons en Fig. 13 leur distribution dans un espace morphologique. L'espace morphologique réduit est obtenu en calculant 4 indicateurs bruts de forme urbaine, qui sont l'index de Moran, la distance moyenne, le niveau de hiérarchie et l'entropie (voir la section 4.1 pour une définition précise et une mise en contexte), et en réduisant la dimension avec une analyse par composantes principales pour laquelle nous gardons les deux premières composantes (92% de variance cumulée). La première mesure un "niveau d'étalement" et d'éclatement, tandis que la

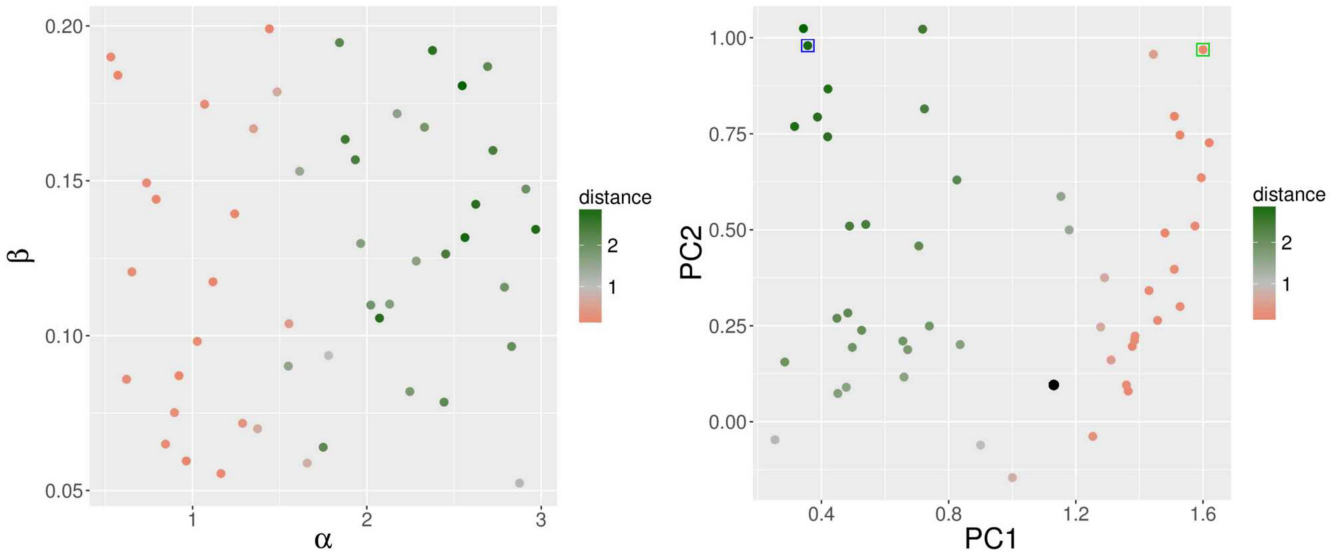

FIGURE 13 : **Distance relative des diagrammes de phase à la référence pour l'ensemble des grilles.** (Gauche) Distance relative comme fonction des meta-paramètres α (force de l'attachement préférentiel) et la diffusion (β, force du processus de diffusion). (Droite) Distance relative comme fonction des deux composantes principales de l'espace morphologique (voir texte). Le point rouge correspond à la configuration spatiale de référence. Les cadres verts et bleu donnent respectivement le premier et le second diagrammes particuliers montrés à la Fig. 14.

seconde mesure l'agrégation.[23] Nous trouvons que les grilles produisant les déviations les plus grandes sont celles avec un faible niveau d'étalement et une forte agrégation. Cela est confirmé par le comportement comme fonction des meta-paramètres, puisque des fortes valeurs de α donnent aussi une forte distance. En terme de processus du modèle, cela montre que les mécanismes de congestion induisent rapidement de plus haut niveaux d'inégalités.

Nous contrôlons à présent la sensibilité en terme de comportement qualitatif des diagrammes de phase. Nous montrons en Fig. 14 les diagrammes pour deux morphologies très opposées en terme d'étalement, mais en contrôlant l'agrégation par la même valeur de PC2. Ceux-ci correspondent au cadres vert et bleu en Fig. 13. Les comportements sont relativement stables pour s_+ variant, ce qui signifie que les agents les plus pauvres ont un rôle déterminant dans les trajectoires. Les deux examples ont non seulement une inégalité de base très distante (le plafond du premier 0.35 est environ le plancher du second 0.3), mais leur comportement qualitatif est également radicalement opposé : la configuration étalée donne des inégalités qui décroissent quand la population décroît et qui décroissent quand la richesse minimale augmente, tandis que la concentrée donne des inégalités augmentant fortement quand la population décroît et aussi

23 Nous avons $\text{PC1} = 0.76 \cdot \text{distance} + 0.60 \cdot \text{entropy} + 0.03 \cdot \text{moran} + 0.24 \cdot \text{slope}$ et $\text{PC2} = -0.26 \cdot \text{distance} + 0.18 \cdot \text{entropy} + 0.91 \cdot \text{moran} + 0.26 \cdot \text{slope}$.

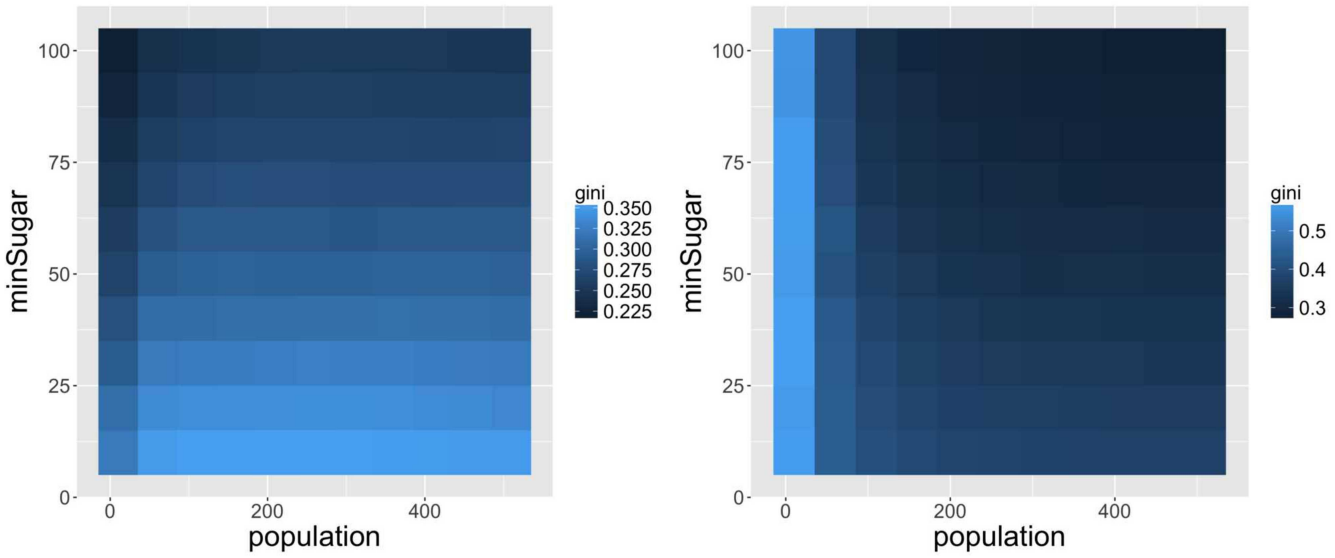

FIGURE 14 : **Exemples de diagrammes de phase.** Nous montrons deux diagrammes bi-dimensionnels sur (P, s_-), obtenus à $s_+ = 110$ fixé. (Gauche) Cadre vert, obtenu avec $\alpha = 0.79$, $n = 2$, $\beta = 0.14$, $N = 157$; (Droite) Cadre bleu, obtenu avec $\alpha = 2.56$, $n = 3$, $\beta = 0.13$, $N = 128$.

décroissantes avec la richesse minimale mais significativement seulement pour des grandes valeurs de population. Le processus est ainsi complètement inversé, ce qui aurait un impact déterminant si l'on essayait de schématiser des politiques à partir du modèle. Cet exemple confirme ainsi l'importance de la sensibilité des modèles de simulation aux conditions spatiales initiales.

⋆ ⋆

⋆

Nous avons vu dans cette section comment nous positionner par rapport à l'usage des modèles de simulation, et plus généralement par rapport au calcul intensif. Nous avons vu revenir de manière récurrente dans les problématiques abordées la question de l'ouverture des pratiques scientifiques.

Nous proposons dans la section suivante d'en détailler un aspect, celle de la reproductibilité, qui en est à la fois une composante mais aussi un produit : simultanément produit et producteur, elle permet une plus grande ouverture et est réciproquement encouragée par les pratiques d'ouverture.

⋆ ⋆

⋆

3.2 REPRODUCTIBILITÉ ET OUVERTURE

La production de connaissance scientifique trouve ses fondements dans la nature cumulative et collective de la recherche, puisque les progrès sont faits lorsque, comme NEWTON l'a bien dit, on "se tient sur les épaules de géants", au sens que l'entreprise scientifique à un temps donné repose sur l'ensemble du travail précédent et qu'aucune avancée ne serait possible sans construire dessus. Cela inclut le développement de nouvelles théories, mais aussi l'extension, le test et la falsification de précédentes : l'avancée dans la construction de la tour signifie aussi la déconstruction de certaines briques obsolètes. Cet aspect de validation par les pairs et de remise en question constante est aussi ce qui légitime la science pour une connaissance plus robuste et un progrès sociétal basés sur une connaissance d'un univers objectif, par rapport aux systèmes dogmatiques qu'ils soient politiques ou religieux [BAIS, 2010].

La reproductibilité semble être de plus en plus pratiquée de manière effective [STODDEN, 2010] et les moyens techniques pour l'achever sont toujours plus développés (comme par exemple les outils pour déposer les données ouvertes, ou pour être transparent dans le processus de recherche comme `git` [RAM, 2013], ou pour intégrer la création de document et l'analyse de données comme `knitr` [XIE, 2013]), au moins dans le champ de la modélisation et de la simulation. Cependant le diable est bien dans les détails et des obstacles jugés dans un premier temps comme mineurs peuvent rapidement devenir un fardeau pour reproduire et utiliser des résultats obtenus dans des recherches précédentes. Nous décrivons deux études de cas où les modèles de simulation sont en apparence hautement reproductibles mais se révèlent vite des puzzles pour lesquels l'équilibre de temps de recherche passe rapidement sous zéro, au sens où essayer d'exploiter leur résultats coûtera plus en temps que de développer entièrement des modèles similaires.

3.2.1 *Explicitation, documentation et implémentation des modèles*

Sur le Besoin d'expliciter le modèle

Un mythe à la vie dure (auquel nous essayons en fait nous-même d'échapper) est que fournir le code source complet et les données seront une condition suffisante pour la reproductibilité, puisque la reproductibilité computationnelle complète implique un environnement similaire ce qui devient vite ardu à produire comme le montrent [HATTON et WARR, 2016]. Pour résoudre ce problème, [HUNG et al., 2016] propose l'utilisation de conteneurs Dockers qui permet de reproduire même le comportement de logiciels avec interface graphique indépendamment de l'environnement. C'est d'ailleurs l'une des direction courantes de développement d'OpenMole, pour simplifier le pa-

ckaging des bibliothèques et des modèles en binaire (voir l'entretien avec R. Reuillon). Dans tous les cas, la reproductibilité a des dimensions supplémentaires, il ne s'agit pas de l'objectif unique qui serait de produire exactement les mêmes graphes et analyses statistiques, en supposant que le code fournit est celui qui a été effectivement utilisé pour produire les résultats donnés. Tout d'abord, ceux-ci doivent être autant que possible indépendants de l'implémentation [Crick, Hall et Ishtiaq, 2017] (c'est-à-dire du langage, des bibliothèques, des choix de structures de données et de type de programmation) pour des motifs clairs de robustesse. Ensuite, en relation avec le point précédent, un des buts de la reproductibilité est la réutilisation des méthodes ou résultats comme base ou modules pour une recherche future (ce qui comprend une implémentation dans un autre langage ou une adaptation de la méthode), au sens que la reproductibilité n'est pas la possibilité stricte de répliquer car elle doit être adaptable [Drummond, 2009].

Notre premier cas d'étude suit exactement ce schéma, puisqu'il a sans aucun doute été conçu pour être partagé avec la communauté et utilisé, s'agissant d'un modèle de simulation fournit avec la plateforme de modélisation agent NetLogo [Wilensky, 1999]. Le modèle est également disponible en ligne [De Leon, Felsen et Wilensky, 2007] et est présenté comme un outil pour simuler les dynamiques socio-économiques des résidents à bas revenus d'une ville au sein d'un environnement urbain synthétique, généré pour ressembler en terme de faits stylisés à la ville réelle de Tijuana au Mexique. Globalement, le modèle fonctionne de la façon suivante : (i) à partir de centres urbains, une distribution d'usage du sol est générée par modélisation procédurale similaire à [Lechner et al., 2006], c'est-à-dire des routes sont générées de proche en proche selon des règles géométriques et de hiérarchie locales, et un usage du sol ainsi qu'une valeur est attribué en fonction des caractéristique de la cellule (distance au centre, à la route) ; (ii) dans cet environnement urbain sont simulées des dynamiques résidentielles de migrants, qui cherchent à optimiser une fonction d'utilité dépendant du coût de la vie et de la configuration des autres migrants. À part fournir le code source, le modèle n'est que peu documenté dans la littérature ou dans les commentaires et la description de l'implémentation. Les commentaires qui suivent sont basés sur l'étude de la partie du modèle simulant la morphogenèse urbaine (initialisation pour la composante "dynamiques résidentielles") comme il s'agit de notre contexte global d'étude. Dans le cadre de cette étude, le code source a été modifié et commenté, dont la dernière version est disponible sur le dépôt du projet[24].

24 À `https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/Reproduction/UrbanSuite`.

FORMALISATION RIGOUREUSE Une partie évidente de la construction d'un modèle est sa formalisation rigoureuse dans un cadre formel distinct du code source. Il n'y a bien sûr aucun langage universel pour le formuler [BANOS, 2013], et de nombreuses possibilités sont offertes par de nombreux champs (e.g. UML, DEVS, formulation mathématique pure), mais l'étape de formalisation précise, qui suit généralement une description plus intuitive donnant les idées et processus dominants, ne peut pas être sautée. On pourrait se dire que le code source y est équivalent, mais ce n'est pas exactement vrai car on pourrait alors ne plus distinguer certains choix d'implémentation de la structure du modèle. Aucun article ni documentation n'accompagne le modèle ici, au delà de la documentation embarquée NetLogo, qui ne décrit que de manière thématique en langage naturel les idées derrière chaque étape sans plus développer et fournit de l'information sur le rôle des différents éléments de l'interface. Comme ces éléments manquent ici, le modèle n'est guère utilisable tel quel. On pourrait nous objecter ici que la partie que nous étudions est une procédure d'initialisation et non le coeur du modèle : nous maintenons que l'ensemble des procédures doit être également documenté et implémenté avec un soin équivalent, ou pointer vers une référence extérieure dans le cas d'utilisation d'un modèle tiers, comme nous le faisons d'ailleurs pour le couplage effectué en 3.1.

Une telle formulation est essentielle pour que le modèle soit compris, reproduit et adapté ; mais elle évite également des biais d'implémentation comme :

- Des éléments architecturaux dangereux : le contexte du monde est une sphère, ce qui n'est pas raisonnable pour ce modèle à l'échelle d'une ville, les mesures de proximité jouant un rôle important dans les processus de production de la forme urbaine. Les agents peuvent passer d'un côté du monde à l'autre dans la représentation euclidienne, ce qui n'est pas acceptable pour une projection en deux dimensions du monde réel. Pour éviter cela, de nombreux tests et fonctions subtiles sont utilisés, incluant des pratiques déconseillées (e.g. mort d'agents basée sur leur position pour les empêcher de sauter).

- Manque de cohérence interne : par exemple la variable de cellule `land-value` (non documentée mais dont l'utilisation se reconstruit par analyse du code) utilisée pour représenter différentes quantités géographiques à différentes étapes du modèle (morphogenèse et dynamiques résidentielles), ce qui devient une incohérence interne quand les deux étapes sont couplées lorsque l'option permettant de faire croître la ville est activée.

- Erreur de code : dans un langage non typé comme NetLogo, le mélange des types peut conduire à des erreurs inattendues à l'execution, ou même des *bugs* non détectables directement

et alors plus dangereux. C'est le cas de la variable de patch `transport` dans le modèle (même si aucune erreur ne survient dans la majorité des configurations depuis l'interface, ce qui est plus dangereux comme le développeur pense que l'implémentation est sûre). De tels problèmes devraient être évités si l'implémentation est faite à partir d'une description exacte du modèle.

IMPLÉMENTATION TRANSPARENTE Une implémentation totalement transparente doit être attendue, incluant une certaine ergonomie dans l'architecture et le code, mais aussi dans l'interface et la description du comportement attendu du modèle.

COMPORTEMENT ATTENDU DU MODÈLE Quelle que soit la définition, un modèle ne peut pas être réduit à sa formulation et/ou implémentation, comme le comportement attendu ou l'utilisation du modèle peuvent être vu comme des parties du modèle lui-même. Dans le cadre du perspectivisme de GIERE [GIERE, 2010c], la définition du modèle inclut le motif de l'utilisation mais aussi l'agent qui vise à l'utiliser. Pour cela une explication minimale du comportement du modèle et une exploration du rôle des paramètres sont fortement recommandés pour diminuer les chances de mauvais usage ou mauvaises interprétations de celui-ci. Cela inclut des graphes simples obtenus immédiatement à l'exécution sur la plateforme NetLogo, mais aussi un calcul d'indicateurs pour évaluer les sorties du modèle. Il peut aussi s'agir de visualisations améliorées pendant l'execution et l'exploration du modèle, comme le montre la figure 15.

Sur le besoin d'exactitude dans l'implémentation du modèle

Des divergences potentielles entre la description du modèle dans un article et les processus effectivement implémentés peut avoir des conséquences graves sur la reproductibilité finale. Le modèle de croissance du réseau routier donné dans [BARTHELEMY et FLAMMINI, 2008] est un exemple d'un tel décalage. Une implémentation stricte des mécanismes du modèle[25] produit des résultats légèrement différents de ceux présentés dans l'article, et comme le code source n'est pas fourni nous devrions tester différentes hypothèses sur des mécanismes possibles ajoutés par le programmeur (qui semble être une règle de connexion aux intersections sous un certain seuil de distance). Des leçons qui peuvent éventuellement être tirées de cet exemple, qui rejoignent partiellement mais complètent celle tirées dans l'étude de cas précédente, sont :

- la nécessité de fournir le code source ;

25 Notre implémentation en NetLogo est disponible à `https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/Reproduction/NWGrowth/LocalDistanceMin`.

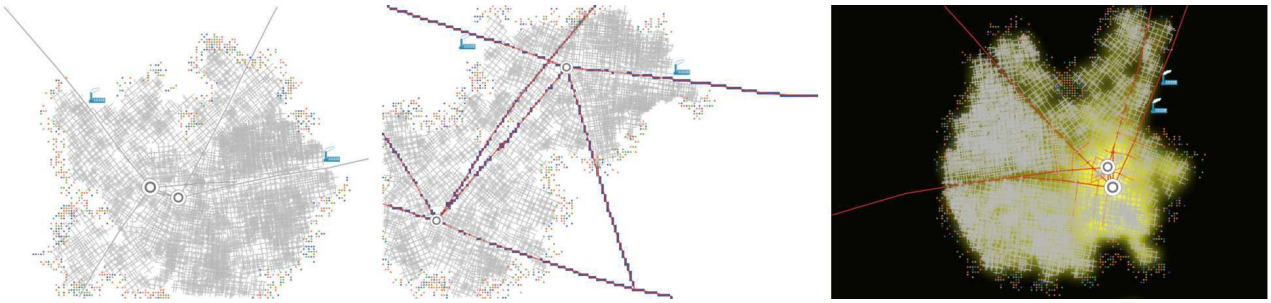

FIGURE 15 : **Exemple d'amélioration simple dans la visualisation qui peut aider à appréhender les mécanismes impliqués par le modèle.** (Gauche) Exemple de sortie originale ; (Centre) Visualisation des routes principales (en rouge) et de l'attribution des patches sous-jacente, qui suggère de possibles biais d'implémentation dans l'utilisation de la trace discrete des routes pour garder trace de leur position ; (Droite) Visualisation des valeurs foncières en utilisant un gradient de couleur plus lisible. Cette étape confirme l'hypothèse, par la forme de la distribution des valeurs, que l'étape de morphogenèse est un détour non-nécessaire pour générer un champ aléatoire pour lequel des simples mécanismes de diffusion devrait fournir des résultats similaires, comme détaillé dans le paragraphe sur l'implémentation. Initialement, l'interface du modèle ne permet pas ces options de visualisation, ces à dire se limite à la première image. On ne peut se rendre compte des processus en jeu pour la morphogenèse, liés aux patches de route et au valeurs foncières se diffusant.

- la nécessité de fournir une description de l'architecture en même temps que le code (si la description du modèle est faite dans un langage trop loin de spécification architecturales) afin d'identifier des biais possibles d'implémentation ;
- la nécessité de procéder à des explorations explicites du modèle et de les détailler, ce qui dans ce cas aurait permis d'identifier de possibles biais d'implémentation.

Rendre le dernier point obligatoire pourrait assurer un risque limité de falsification puisqu'il est généralement plus compliqué de falsifier des résultats d'exploration plutôt que d'explorer effectivement le modèle. On pourrait imaginer une expérience pour tester le comportement général d'un sous-ensemble de la communauté scientifique au regard de la reproductibilité, qui consisterait en l'écriture d'un faux article de modélisation dans l'esprit de [ZILSEL, 2015], dans lesquels des résultats opposés aux résultats effectifs d'un modèle donné seraient fournis, sans fournir l'implémentation du modèle. Un premier test serait de tester l'acceptation d'un article clairement non reproductible dans divers journaux, si possible avec un contrôle sur les éléments textuels (par exemple en utilisant ou non des "buzzwords" chers au journal). Selon les résultats, une expérience plus poussée serait de fournir l'implémentation open source mais toujours avec des résultats modifiés plus ou moins fortement, afin de tester si les reviewers essayent effectivement de reproduire les résultats quand ils demandent le code (dans des capacités de calcul limitées bien sûr, le calcul intensif n'étant pas encore largement disponibles en sciences

sociales). Notre intuition est que les résultats obtenus seraient fortement négatifs, vu les difficultés rencontrées par une exigence de discipline de reproduction indépendante lors de nombreuses relectures, même pour des revues faisant de la reproductibilité une condition *sine qua non* de la publication, les auteurs trouvant des astuces pour se dérober aux contraintes (postuler que des données de simulation ne sont pas des données, ne fournir qu'une version agrégée inutile du jeu de données utilisées, etc. ; nous reviendrons sur le rôle des données plus loin).

Exploration interactive et production des résultats

L'usage d'applications interactives pour la fouille de données a des avantages non discutables, tel qu'une familiarisation avec la structure des données par une vue d'ensemble qui serait beaucoup plus laborieuse voire impossible autrement. C'est la même idée sous-jacente qui justifie l'interactivité pour l'exploration préliminaire des modèles multi-agents intégrée à des plateformes comme Netlogo [Wilensky, 1999] ou Gamma [Drogoul et al., 2013]. Un objectif similaire est implicite dans [Rey-Coyrehourcq, 2015], c'est-à-dire une intégration complète de l'exploration fine des modèles et de la production des graphes de sortie ainsi que leur exploration interactive. Comme le rappelle Romain Reuillon (Entretien du 11/04/2017, voir D.3), la plateforme OpenMole qui devait accueillir cette couche supplémentaire était à ses débuts à l'époque et ne l'est toujours pas aujourd'hui, puisque l'état de l'art de telles pratiques est en pleine construction et bouleversements réguliers [Holzinger, Dehmer et Jurisica, 2014].

Des difficultés au regard de la reproductibilité, qui nous concernent particulièrement ici, sont récurrentes et loin d'être résolues. En effet, il faut bien situer la position de ces outils et méthodes comme une aide cognitive préliminaire[26], mais peu souvent comme permettant la production de résultats finaux : lorsque les paramètres ou dimensions se multiplient, l'export d'un graphe est bien souvent déconnecté de l'information complète ayant conduit à sa production. De la même manière, l'utilisation de notebooks intégrés tel Jupyter, permettant d'intégrer analyses et rédaction du compte-rendu, peut devenir dangereux car on peut justement revenir sur un script, tester différentes valeurs d'un paramètre, et perdre les valeurs qui avaient produit un graphe donné. L'utilisation de versioning peut être une solution partielle mais souvent lourde.

Dans l'idéal, tout logiciel interactif permettant l'export de résultats devrait en même temps exporter un script ou une description exacte et utilisable permettant d'arriver exactement à ce point à partir des données brutes. La plupart des applications d'exploration interactives de données spatio-temporelles sont à ce regard relativement imma-

26 Que nous ne jugeons pas superficielle puisque nous les mobilisons au moins deux fois par la suite, voir ci-dessous ainsi que C.1.

tures scientifiquement, car même dans le cas où elles sont totalement honnêtes et transparentes sur les analyses présentées à l'utilisateur, ce qui n'est malheureusement pas la règle, les tâtonnements d'exploration progressive ne sont pas reproductibles et la méthode d'extraction de caractéristiques est ainsi relativement aléatoire. En poussant le raisonnement, leur utilisation révélerait plutôt l'aveu d'une faiblesse d'un manque de méthodes systématiques accompagnant la découverte de motifs dans des données spatio-temporelles complexes de manière efficace.

Par un plaidoyer visionnaire, Banos avait déjà mis en garde contre "les dangers de la jungle" des données dans [Banos, 2001], quand il souligne très justement que l'exploration interactive doit nécessairement se doubler d'indicateurs locaux adaptés, mais surtout d'outils d'exploration automatisés et de critère d'évaluation des choix faits et des motifs découverts par l'utilisateur. Nous revenons encore à l'idée d'une plateforme intégrée dont OpenMole pourrait être un précurseur. La combinaison des capacités cognitives humaines au traitement machine, notamment pour des problèmes de vision par ordinateur, ouvre des possibilités de découvertes inédites, encore plus via une utilisation collective comme en témoigne le Galaxy Zoo [Raddick et al., 2010][27]. Les résultats d'un crowdsourcing de la cognition humaine peuvent rivaliser avec les techniques automatiques les plus avancées comme le montre [Koch et Stisen, 2017] pour l'exemple de la comparaison de cartes spatiales.

Ces possibilités ne doivent cependant pas être sur-estimées ou utilisées à mauvais escient, et les questions d'intégration efficiente homme-machine sont d'ailleurs totalement ouvertes. Dans le domaine de la visualisation de l'information géographique, [Pfaender, 2009] introduit une sémiologie spécifique visant à favoriser l'exploration de grands jeux de données hétérogènes, et l'expérimente sur une application spécifique : il s'agit d'une avancée considérable vers une plateforme intégrée et une exploration interactive saine et reproductible, les directions d'exploration répondant à des modèles basés sur les sciences cognitives.

Enfin, le rôle de l'interactivité dans la communication et la vulgarisation scientifiques est exploré par l'Annexe C.7, qui suggère la mise en place de jeux, notamment un jeu informatique interactif, pour faciliter la transmission de concepts scientifiques au public. cela nous montre que le développement de ces pratiques innovantes dépasse largement le seul cadre de l'analyse de données.

27 Le principe rejoint celui de *citizen science*, en faisant participer des volontaires hors de la communauté scientifique à des tâches requérant cognition mais pas de connaissances scientifique : la classification d'images, dans le but d'entraîner des algorithmes supervisés, est l'exemple initial du Galaxy Zoo pour la forme des galaxies.

Mise en application

Encore une fois, la reproductibilité et la transparence sont des éléments essentiels incontournables de la science contemporaine, liés aux pratiques de science ouverte et d'accès ouvert. Beaucoup d'exemples (voir un récent en économie expérimentale dans [Camerer et al., 2016]) dans diverses disciplines montrent le manque de reproductibilité des résultats des expériences, alors que celle-ci doit pouvoir conduire à une falsification ou à une confirmation de ces résultats. La falsification est une pratique coûteuse car demandant un certain investissement au détriment de sa propre recherche [Chavalarias et al., 2005]. Elle pourrait ainsi être rendue plus efficiente grâce à une transparence augmentée. Des outils spécialement dédiés à une reproductibilité directe, souvent permise par l'ouverture, devraient accroître la performance globale de la science. Mais l'accès ouvert a des impacts bien plus larges que la science elle-même : [Teplitskiy, Lu et Duede, 2017] montrent un transfert des connaissances scientifiques accru vers la société dans le cas d'articles ouverts, notamment par des intermédiaires comme Wikipedia.

Le développement et la systématisation de standards et de bonnes pratiques, de manière conjointe sur les différentes problématiques évoquées, est une condition nécessaire à une rigueur scientifique qui devrait être uniforme au travers de l'ensemble des disciplines existantes. Nous construisons par exemple des outils facilitant le flot de production scientifique, ceux-ci étant détaillés en Appendice E.3. Par exemple, pour les sciences computationnelles, on a déjà évoqué les potentialités de l'utilisation de `git` qui s'étendent en fait sans contrainte de disciplines ni de types de recherche si les bonnes adaptations sont introduites. Le suivi précis de l'ensemble des étapes d'un projet, gardé en historique offrant la possibilité de revenir à n'importe laquelle à tout moment, mais aussi de travailler de façon collaborative, plus ou moins parallèlement selon les besoins en utilisant les branches, est un exemple de service fourni par cet outil. Un exemple de bonnes pratiques d'utilisation est donné par [Perez-Riverol et al., 2016].

Plus généralement, les sciences computationnelles nécessitent l'adoption de certains standards et pratiques pour assurer une bonne reproductibilité, et ceux-ci restent majoritairement à développer : [Wilson et al., 2017] donne des premières pistes. Concernant la qualité des données, de nombreux efforts sont faits pour introduire des cadres de standardisation des données : par exemple [Veiga et al., 2017] décrit un cadre conceptuel visant à guider la résolution de problème récurrent liés à la qualité des données de biodiversité (comme par exemple évaluer des mesures jugeant de l'usage possible d'un jeu de données pour un problème donné). De nouvelles perspectives s'ouvrent pour des futurs cadres de traitement de données intrinsèquement ouverts

et reproductibles, avec le développement de nouvelles techniques comme le *blockchain*[28], comme proposé par [Furlanello et al., 2017].

3.2.2 *Ouverture des données*

L'accès aux données est également un point crucial pour la reproductibilité, et sans nous y attarder car cela impliquerait des développements sur la définition, la philosophie, le droit des données etc. qui sont des sujets de recherche en eux-mêmes, nous donnons des perspectives sur les opportunités offertes par une ouverture systématique des données en recherche. En géographie, les *data paper* sont une pratique inexistante, et la règle est plutôt de garder la main jalousement sur un jeu produit, capitalisant sur le fait d'être le seul à y avoir accès[29].

Il est évident que la qualité et quantité des connaissances produites sera nécessairement plus grande si un jeu de données est publiquement ouvert, puisqu'au moins la même chose sera obtenue, et on peut s'attendre à une prise en main par d'autres domaines, d'autres méthodes, et donc à une plus grande richesse[30].

La fermeture induira plutôt des effets négatifs, comme par exemple du temps perdu à recoder un base vectorielle donnée uniquement sous forme de carte dans un article. L'argument du temps passé comme justification à la fermeture est absurde, puisqu'au contraire, en voyant les données comme une composante à part entière de la connaissance (voir le cadre de connaissances en 8.3), le temps passé doit impliquer plus de citations, donc plus d'utilisation, ce qui passe nécessairement par l'ouverture pour des données. De même, quelle logique, sinon la même absurde de propriété des connaissances, pousse les géographes à insérer un copyright sur l'ensemble de leurs cartes mais aussi leurs figures, jusqu'à un copyright pour un simple histogramme qui s'en serait bien passé si on avait pu l'interroger, honnête de simplicité ?

28 Le *blockchain* consiste en la distribution d'un graphe de transactions entre utilisateurs, celles-ci étant validées (dans le cadre historique classique de type *proof-of-work*) par la résolution de problèmes cryptographiques inverses par force brute, par des agents appelés mineurs, essentiels à la robustesse de l'écosystème.

29 Il n'existe à notre connaissance pas de travail quantifiant la proportion de données ouvertes sur l'ensemble des données produites en géographie. Cela pourrait être l'objet d'un travail d'épistémologie quantitative appliquant des techniques similaires à celles développées en chapitre 2. La difficulté à trouver des données ouvertes, comparée à la fréquence des publications dans les domaines concernés, suggère une validité au moins qualitative de ce fait.

30 Il est possible d'argumenter que le système de production scientifique est complexe, et qu'une monétarisation, compétition ou privatisation accrue de la recherche peut faire partie d'un écosystème de recherche dont les sorties pourront être jugées de qualité selon les indicateurs choisis. Ces considérations sont pertinentes, mais hors de notre portée puisque relevant d'un travail en anthropologie et sociologie des sciences. Nous postulons ici ce principe, et le considérons comme une position scientifique subjective.

L'expérience d'évaluation d'articles nous induit à réellement nous inquiéter sur la valeur donnée à l'ouverture des données par les auteurs : au bout d'une dizaine d'articles, incluant des journaux affichant comme priorité et pré-requis l'ouverture totale des données et modèles, dont un seul est seulement partiellement ouvert et l'ensemble des autres implique de croire sur parole les résultats présentés (alors qu'un des but de la revue est de contourner les biais cognitifs qu'un ou des humains ont forcément par une validation croisée qui doit se faire sur les résultats bruts et non des interprétations contenant ces biais), il est difficile de croire que des mutations profondes des pratiques ne sont pas nécessaires.

Mais en suivant l'adage de Framasoft[31], "la route est longue mais la voie est libre", les perspectives sont nombreuses pour une évolution dont la lenteur n'est pas inéluctable. Le journal Cybergéo, pionnier des pratiques d'ouverture en sciences sociales (première revue entièrement électronique, première revue à lancer une rubrique de *model papers*), lance en 2017 une rubrique *data papers*[32] visant à inciter le développement du partage de données et de l'ouverture en géographie.

Il reste des zones grises sur lesquelles il est impossible aujourd'hui d'avoir des perspectives, notamment le droit des données. Nous avons un exemple dans les analyses que nous développerons : les données bibliographiques sont obtenues au prix d'une guerre de blocage par Google et un effort technique considérable pour la gagner (voir 2.2 et B.6).

L'ouverture implique un engagement qui fait résolument partie de nos positionnements. C'est la même idée qui soutient la construction de l'application `CybergeoNetworks`[33], qui couple les outils présentés en 2.2 avec d'autres approches complémentaires d'analyse de corpus, dans le but d'encourager la réflexivité scientifique, et de mettre cet outil ouvert à la disposition d'éditeurs indépendants, pour s'émanciper de la nouvelle main mise des géants de l'édition qui à la recherche d'un nouveau modèle pour sécuriser leur profits parient sur la vente de méta-contenu et de son analyse. Heureusement, la récente loi numérique en France a gagné le bras de fer contre leur revendication d'un droit exclusif sur la fouille de texte complets.

3.2.3 *Illustration par une étude empirique*

Nous proposons à présent de développer un exemple concret d'étude empirique illustrant les derniers points relevés ci-dessus et nous permettant une entrée progressive dans notre problématique. Dans le cas du trafic routier en Ile-de-France, nous menons une collecte d'un

31 Réseau pour la promotion du logiciel libre, `https://framasoft.org/`

32 Dont l'index est disponible à `https://cybergeo.revues.org/28545`. Le premier article est [Swerts, 2017], que nous utilisons d'ailleurs en 7.3.

33 Dont la démarche et le contexte sont détaillés en Annexe C.4. Elle est disponible en ligne à `http://shiny.parisgeo.cnrs.fr/CybergeoNetworks`.

jeu de données là où il n'existe pas de source ouverte. Nous mettons également en place une application permettant son exploration interactive.

Nous avons développé en 1.1 le concept de mobilité quotidienne comme jouant un rôle clé dans les processus d'interaction entre réseaux de transport et territoires, à une échelle que nous avons désignée par microscopique. Il est de plus candidat à la mobilisation de dynamiques co-évolutives, comme le suggère l'effet des localisations sur la congestion et réciproquement.

Ici, la mobilité sera captée par le flux de trafic, et la co-évolution s'opère entre propriétés du réseau (congestion) et localisation des agents. Nous nous intéresserons plus particulièrement à l'équilibre hypothétique des flux de trafic, répondant indirectement à des problématiques que nous détaillons ci-dessous.

Contexte

La modélisation du trafic a été largement étudiée depuis les travaux séminaux de Wardrop ([Wardrop, 1952]) : les enjeux économiques et techniques justifient le besoin d'une compréhension fine des mécanismes régissant les flux de trafic à différentes échelles. Des approches aux objectifs différents coexistent aujourd'hui, parmi lesquels on trouve par exemple les modèles dynamiques de micro-simulation, généralement opposés aux techniques se basant sur l'équilibre.

Tandis que la validité des modèles microscopiques a été étudiée de façon conséquente et leur application souvent questionnée, la littérature est relativement pauvre en études empiriques testant l'hypothèse d'équilibre stationnaire du cadre de l'Equilibre Utilisateur Statique (EUS).

De nombreux développements plus précis dans les hypothèses de modélisation ont été documentés dans la littérature, tels l'Equilibre Utilisateur Dynamique Stochastique (EUDS) (voir pour une description par example [Han, 2003]). À un niveau intermédiaire entre les cadres statiques et stochastiques se trouve l'Equilibre Utilisateur Stochastique Restreint, pour lequel les choix d'itinéraire des utilisateurs sont contraints à un ensemble d'alternatives réalistes ([Rasmussen et al., 2015]).

D'autres extensions prenant en compte le comportement de l'utilisateur via des modèles de choix ont été proposé plus récemment, comme [Zhang, Mahmassani et Lu, 2013] qui inclut à la fois l'influence de la tarification routière et de la congestion sur le choix avec un modèle Probit. La relaxation d'autres hypothèses restrictives comme la maximisation pure de l'utilité par l'utilisateur ont aussi été introduites, tels l'Equilibre Utilisateur Borné décrit par [Mahmassani et Chang, 1987]. Dans ce cadre, l'utilisateur est satisfait si sa fonction d'utilité rentre dans une plage de valeurs tolérables, et l'équilibre est achevé lorsque chaque utilisateur est satisfait. Les dynamiques

résultantes sont plus complexes comme révélé par l'existence d'équilibres multiples, et permet de rendre compte de faits stylisés spécifiques comme des évolutions irréversibles du réseau comme développé par [Guo et Liu, 2011].

D'autres modèles d'attribution de trafic inspirés d'autres domaines ont également été plus récemment proposés : dans [Puzis et al., 2013], une définition étendue de la centralité de chemin qui combine linéairement le centralité des flots non-contraints avec une centralité pondérée par le temps de parcours permet d'obtenir une forte corrélation avec les flux de trafic effectifs, fournissant ainsi un modèle d'attribution de trafic. Cela fournit également des applications pratiques comme l'optimisation de la distribution spatiale des capteurs de trafic.

Malgré ces nombreux développements, de nombreuses études et applications concrètes se basent sur l'Equilibre Utilisateur Statique. La région parisienne utilise par exemple un modèle statique (MODUS) pour gérer et planifier le trafic. [Leurent et Boujnah, 2014] introduit un modèle statique de flots qui inclut les recherches locales et le choix du parking : dans ce cas particulier à de si faibles échelles, la stationnarité de la distribution des flux a encore moins de chances d'être une réalité. Un example d'exploration empirique des hypothèses classiques est donné par [Zhu et Levinson, 2015], pour lequel les choix d'itinéraires révélés sont étudiés. Les conclusions questionnent le "premier principe de Wardrop" qui postule que les utilisateurs choisissent parmi un ensemble d'alternatives parfaitement connu.

Dans le même esprit, nous proposons d'étudier l'existence empirique de l'équilibre statique. Plus précisément, l'EUS suppose une distribution stationnaire des flux sur l'ensemble du réseau. Cette hypothèse reste valable dans le cas d'une stationnarité locale, tant que l'échelle temporelle d'évolution des paramètres est considérablement plus grande que les échelles typiques de voyage. Le second cas qui est plus plausible et de plus compatible avec les cadres théoriques dynamiques est testé ici. L'objectif de ce développement est ainsi d'étudier à une grande échelle les relations entre réseaux et territoires, par l'intermédiaire des flux de trafic qui sont portés par le réseau mais générés par les motifs territoriaux.

Dans un premier temps, la procédure de collection de données ainsi que le jeu de données sont décrits ; nous présentons ensuite une application interactive pour l'exploration du jeu de données, dans le but de fournir une intuition sur les motifs présents ; puis nous donnons divers résultats d'analyses quantitatives allant dans le sens d'indices convergents pour une non-stationnarité des flux de trafic.

Jeu de données

CONSTRUCTION DU JEU DE DONNÉES Nous proposons de travailler sur l'étude de cas de la métropole parisienne. Un jeu de données ouvert a été construit, comprenant les liens autoroutiers du coeur urbain dense[34], par collecte des données publiques en temps réel des temps de parcours (disponible sur `www.sytadin.fr`). Comme rappelé par [BOUTEILLER et BERJOAN, 2013], la disponibilité de jeux de données ouverts pour les transports est loin d'être la règle, et nous contribuons ainsi à une ouverture par la construction de notre jeu de données. La procédure de collecte de données consiste en les points suivants, exécutés toutes les deux minutes par un script `python` :

- récupération de la page web brute donnant les informations de trafic
- parsing du code html afin de récupérer les identifiants des liens de trafic et les temps de parcours correspondants
- insertion des liens dans une base `sqlite` avec le temps courant.

Le script automatisé de collection des données continue d'enrichir la base au fur et à mesure du temps, permettant des développements futurs de ce travail sur un jeu de données plus large, et une réutilisation potentielle pour des travaux scientifiques ou opérationnels. La dernière version du jeu de données au format sqlite est disponible en ligne sous une Licence *Creative Commons*[35].

DESCRIPTION DES DONNÉES Une granularité de deux minutes a été obtenue pour une période de trois mois (de février 2016 à avril 2016 inclus)[36]. La granularité spatiale (la distance moyenne entre les centroïdes des liens) est en moyenne de 10km, les temps de trajet étant fournis pour les liens majeurs. Le jeu de données contient 101 liens. La variable brute utilisée est le temps de trajet effectif, à partir duquel il est possible de construire la vitesse de trajet et la vitesse relative de trajet, définie comme le rapport entre temps de trajet optimal (temps de trajet sans congestion, pris comme le temps minimal sur l'ensemble des pas de temps) et le temps de trajet effectif. La congestion est calculée par inversion d'une fonction BPR[37] simple avec exposant 1 comme il est fait par [BARTHELEMY, 2016a], i.e. en prenant $c_i = 1 - \frac{t_{i,min}}{t_i}$ avec t_i temps de trajet effectif dans le lien i et $t_{i,min}$ temps de trajet minimal.

34 Majoritairement Paris et les départements de la petite couronne.

35 Sur le dataverse au lien `http://dx.doi.org/10.7910/DVN/X22ODA`.

36 Comme nous allons travailler à l'échelle temporelle intra-journalière, nous n'avons pas besoin d'un jeu de données plus étendu dans le temps pour avoir des conclusions significative comme nous le verrons par la suite.

37 Il s'agit d'une fonction permettant de relier vitesse à congestion dans un lien, largement utilisée en ingénierie des transports [BRANSTON, 1976].

Analyse des motifs de trafic

VISUALISATION DES MOTIFS SPATIO-TEMPORELS DE CONGESTION Notre approche étant entièrement empirique, une bonne connaissance des motifs existants pour les variables de trafic, en particulier de leur variations spatio-temporelles, est crucial pour guider toute analyse quantitative. En s'inspirant de la littérature étudiant la validation empirique de modèles, plus précisément les techniques de *modélisation orientée-motifs* introduites par [GRIMM et al., 2005], nous nous intéressons aux motifs macroscopiques, par exemple les corrélations, à des échelles temporelles et spatiales données : d'une manière équivalente aux faits stylisés qui sont dans cette approche extraits d'un système avant de tenter de le modéliser, nous devons explorer les données de manière interactive dans le temps et l'espace afin d'identifier des motifs pertinents et les échelles associées.

Une application web interactive a ainsi été implémentée pour explorer les données, à l'aide des packages `R shiny` et `leaflet`[38]. L'application permet une visualisation dynamique des motifs de congestion sur l'ensemble du réseau ou dans une zone particulière grace au zoom. L'application est accessible en ligne à l'adresse `http://shiny.parisgeo.cnrs.fr/transportation`. La Figure 16 présente une capture d'écran de l'interface.

La conclusion majeure de l'exploration interactive des données est qu'une grande hétérogénéité spatiale et temporelle est la règle. Le motif temporel le plus récurrent, la périodicité journalière des heures de pointe, est perturbée pour une proportion non négligeable de jours. En première approximation, les heures creuses peuvent être approchées par une distribution localement stationnaire des flux, tandis que la courte durée des heures de pointe suggère un système non-stationnaire sur ces périodes. Concernant l'espace, aucun motif spatial particulier n'émerge clairement. Cela signifie que dans le cas d'une validité de l'équilibre utilisateur statique, les méta-paramètres régissant son établissement doivent varier à des échelles temporelles plus courtes qu'un jour.

Nous postulons au contraire que le système de trafic est loin de l'équilibre, en particulier pendant les heures de pointe pendant lesquelles des transitions de phase critiques à l'origine des embouteillages émergent.

VARIABILITÉ SPATIO-TEMPORELLE DES TRAJETS A la suite de l'exploration interactive des données, nous proposons de quantifier la variabilité spatiale des motifs de congestion pour valider ou invalider l'intuition que si l'équilibre existe par rapport au temps, il est fortement dépendant de l'espace et localisé. La variabilité spatio-temporelle des plus courts chemins de trajet est une première façon

38 Le code source de l'application et des analyses est disponible sur le dépôt ouvert du projet à `https://github.com/JusteRaimbault/TransportationEquilibrium`.

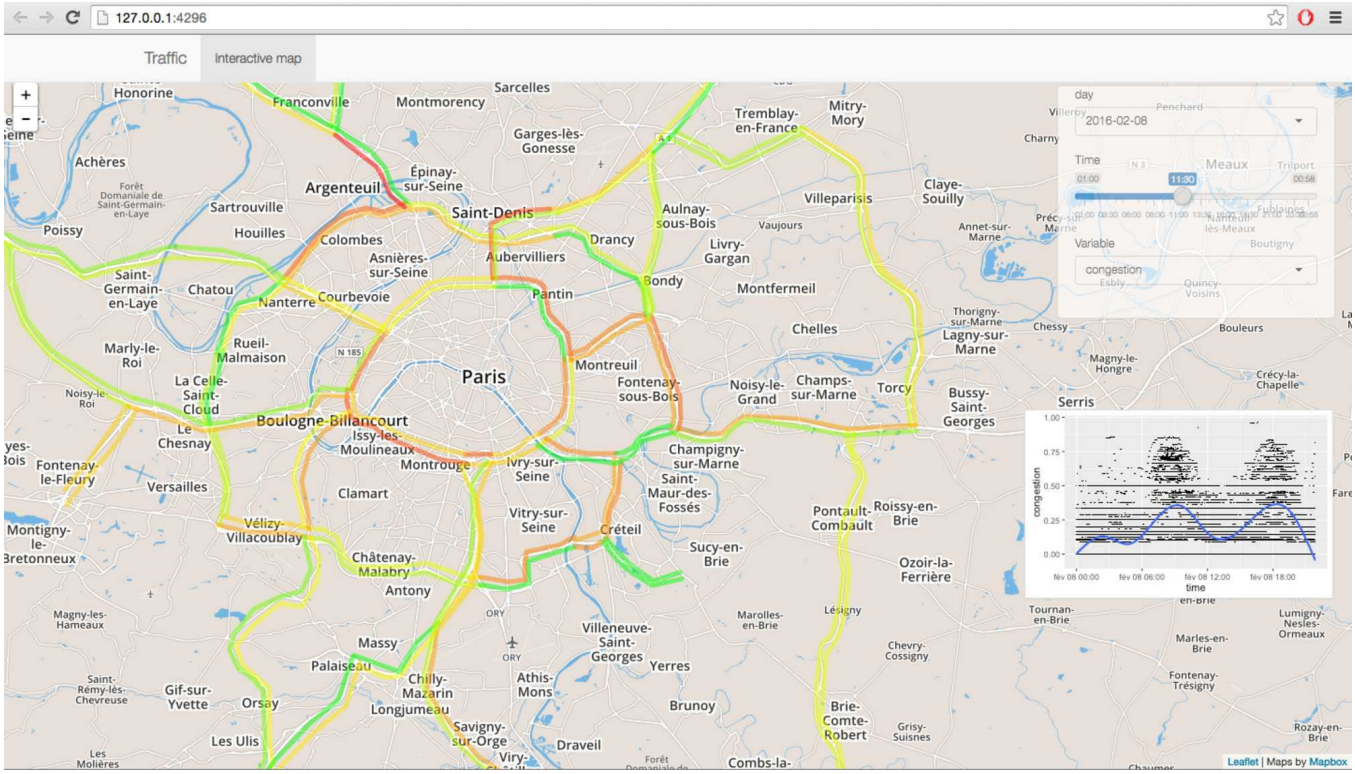

FIGURE 16 : **Capture de l'application web.** Nous avons développé celle-ci pour permettre l'exploration spatio-temporelle des données de trafic pour la région Parisienne. Il est possible de choisir date et heure (précision de 15min sur un mois, réduite par rapport au jeu de données initial pour des raisons de performance). Le graphe en insert résume les motifs de congestion pour la journée courante, en donnant en fonction du temps l'ensemble des valeurs (points noirs) et leur lissage (courbe bleue).

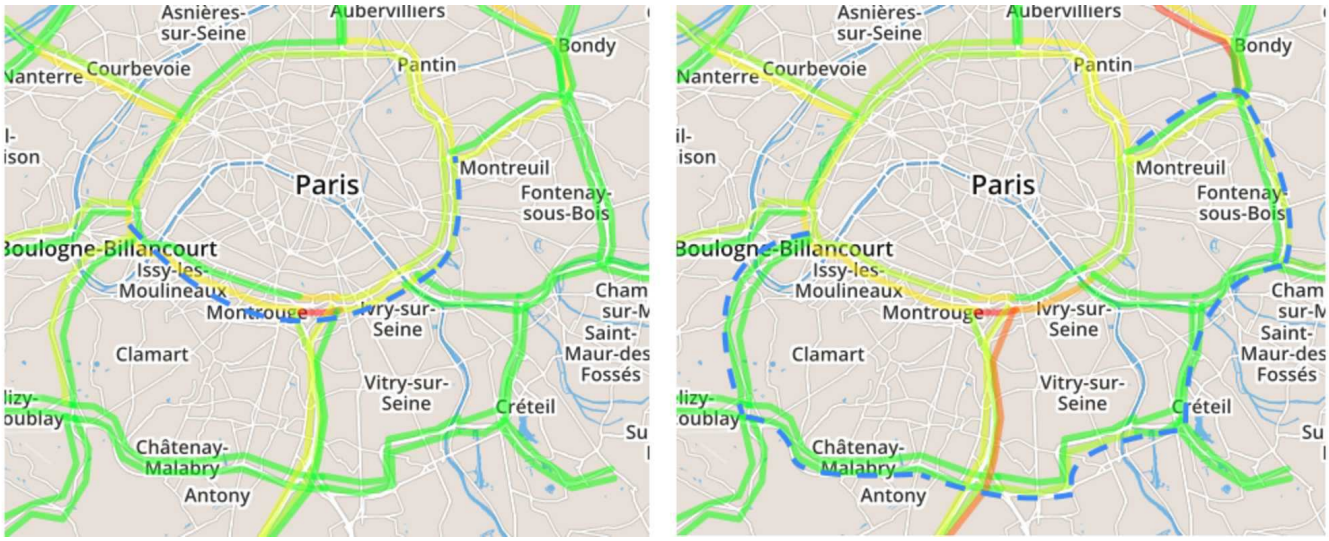

FIGURE 17 : **Variabilité spatiale d'un plus court chemin en temps de trajet.** Le trajet du plus court chemin est donné en pointillé bleu. Dans un intervalle de seulement 10 minutes, entre le 11/02/2016 00 :06 (à gauche) et le 11/02/2016 00 :16 (à droite), le plus court chemin entre Porte d'Auteuil à l'ouest et Porte de Bagnolet à l'est, augmente en distance effective de $\simeq$ 37km (avec une augmentation du temps de trajet de seulement 6 minutes), à cause d'une forte perturbation sur le périphérique parisien.

d'étudier la stationnarité des flux d'un point de vue de théorie des jeux. En effet, l'Equilibre Utilisateur Statique est la distribution stationnaire des flux sous laquelle aucun utilisateur ne peut augmenter son temps de trajet en changeant son itinéraire. Une forte variabilité spatiale des plus courts chemins sur de courtes échelles spatiales révèle ainsi une non-stationnarité, puisqu'un même utilisateur prendra un chemin complètement différent après un court laps de temps et ne contribuera plus au même flux que précédemment. Une telle variabilité est en effet observée sur un nombre non-négligeable de chemins pour chaque jour du jeu de données. La figure 17 montre un exemple de variation spatiale extrême d'un trajet pour une paire Origine-Destination particulière.

L'exploration systématique de la variabilité du temps de trajet sur l'ensemble du jeu de données, et des distances de trajet associées, confirme, comme présenté en figure , que la variation absolue du temps de trajet présente fréquemment une forte variation de son maximum sur l'ensemble des paires O-D, jusqu'à 25 minutes avec une moyenne temporelle locale autour de 10 minutes. La variabilité spatiale correspondante entraine des détours allant jusqu'à 35km.

STABILITÉ DES MESURES DE RÉSEAU La variabilité des trajectoires potentielles observée dans la section précédente peut être confirmée par l'étude de la variabilité des propriétés du réseau. En particulier, les mesures topologiques de réseau capturent les motifs globaux dans un réseau de transport. Les mesures de centralité et de connectivité

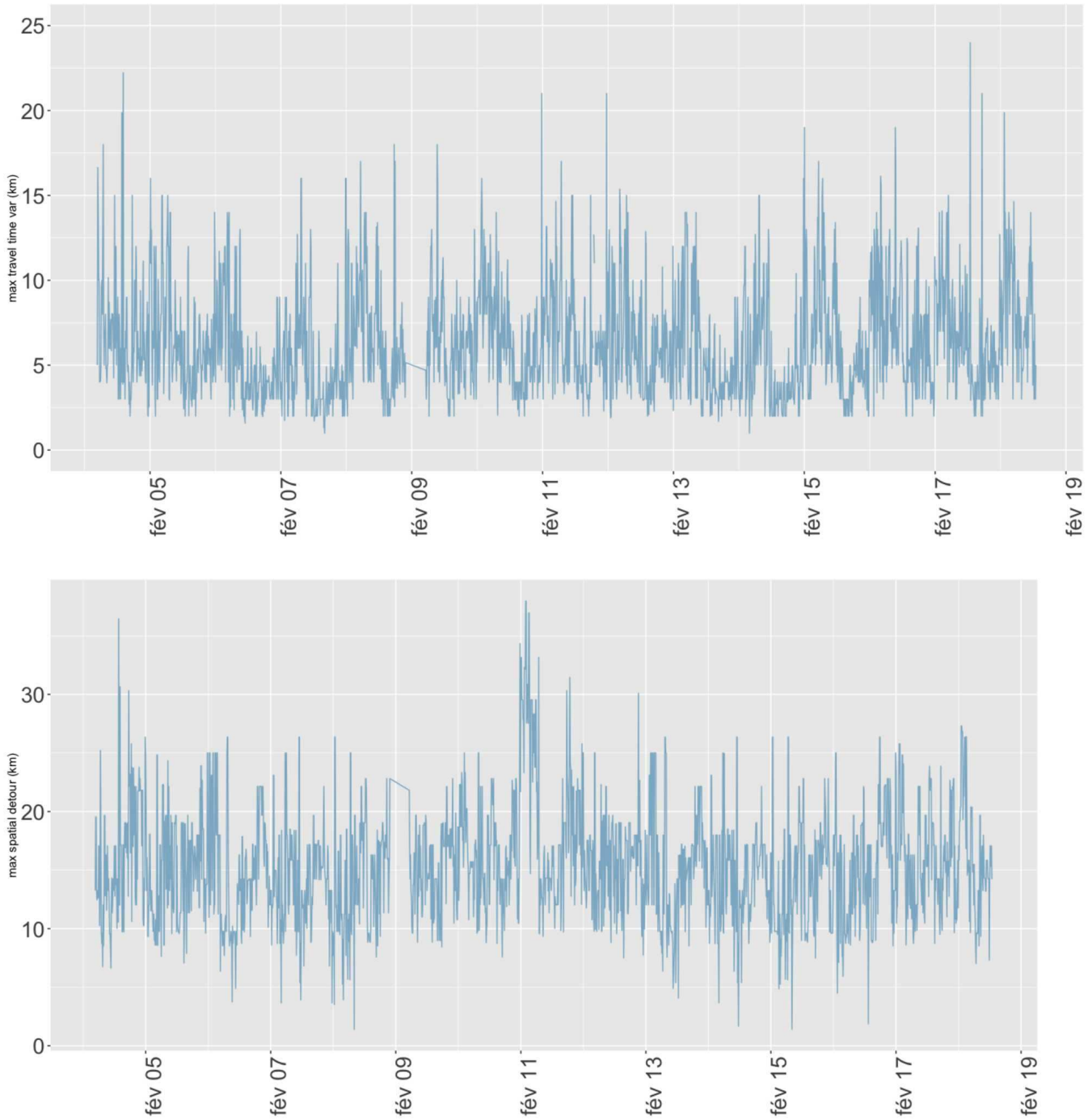

FIGURE 18 : **Variabilité des temps de trajet.** Variabilité maximale du temps de trajet (*Haut*) en minutes et de la distance de trajet correspondante (*Bas*) pour un échantillon de deux semaines. Le graphe représente le maximum sur l'ensemble des paires Origine-Destination de la variabilité absolue entre deux pas de temps consécutifs. Les heures de pointe induisent une forte variabilité du temps de trajet, allant jusqu'à 25 minutes et une variabilité de distance jusqu'à 35km.

des noeuds sont des indicateurs classiques pour la description des réseaux de transport comme rappelé par [Bavoux et al., 2005]. La littérature en transports a développé des mesures de réseau élaborées et opérationnelles, comme des mesures de robustesse pour identifier les liens critiques et mesurer la résilience globale du réseau aux perturbations (un exemple parmi d'autres est l'indice de *Robustesse du Réseau Effective* introduit dans [Sullivan et al., 2010]).

Plus précisément, nous étudions la centralité de chemin du réseau de transport, défini pour un noeud comme le nombre de plus courts chemins passant par celui-ci, i.e. par l'équation

$$b_i = \frac{1}{N(N-1)} \cdot \sum_{o \neq d \in V} \mathbb{1}_{i \in p(o \to d)} \tag{2}$$

où V est l'ensemble des sommets du réseau de taille N, et $p(o \to d)$ est l'ensemble des noeuds sur le plus court chemin entre les sommets o et d (le plus court chemin étant calculé avec le temps de trajet effectif). Cette mesure de centralité est plus adaptée que d'autres dans notre cas, comme la centralité de proximité qui n'inclut pas la congestion potentielle comme la centralité de chemin.

Nous montrons en Fig. 19 la variation relative absolue du maximum de la centralité de chemin, pour la même fenêtre temporelle que les indicateurs empiriques précédents. Plus précisément, elle est définie par :

$$\Delta b(t) = \frac{|\max_i(b_i(t + \Delta t)) - \max_i(b_i(t))|}{\max_i(b_i(t))} \tag{3}$$

où Δt est le pas de temps du jeu de données (la plus petite fenêtre temporelle sur laquelle une variabilité peut être capturée). Cette variation relative absolue a une signification directe : une variation de 20% (qui est atteinte un nombre significatif de fois comme montré en Figure 19) implique dans le cas d'une variation négative, qu'au moins cette proportion de trajectoires potentielles ont changé et que la potentielle congestion locale a décru de la même proportion. Dans le cas d'une variation positive, un seul noeud a capturé au moins 20% des trajets.

Sous l'hypothèse (qu'on ne tente pas de vérifier ici et qu'on peut également supposer non vérifiée comme montré par [Zhu et Levinson, 2015], mais que l'on utilise comme un outil pour donner une intuition sur la signification concrète de la variabilité de la centralité) que les utilisateurs choisissent rationnellement le plus court chemin, et supposant que la majorité des trajets est réalisée, une telle variation de la centralité implique une variation similaire dans les flux effectifs, conduisant à la conclusion qu'ils ne peuvent être stationnaires ni dans le temps (au moins sur une échelle plus grande que Δt) ni dans l'espace.

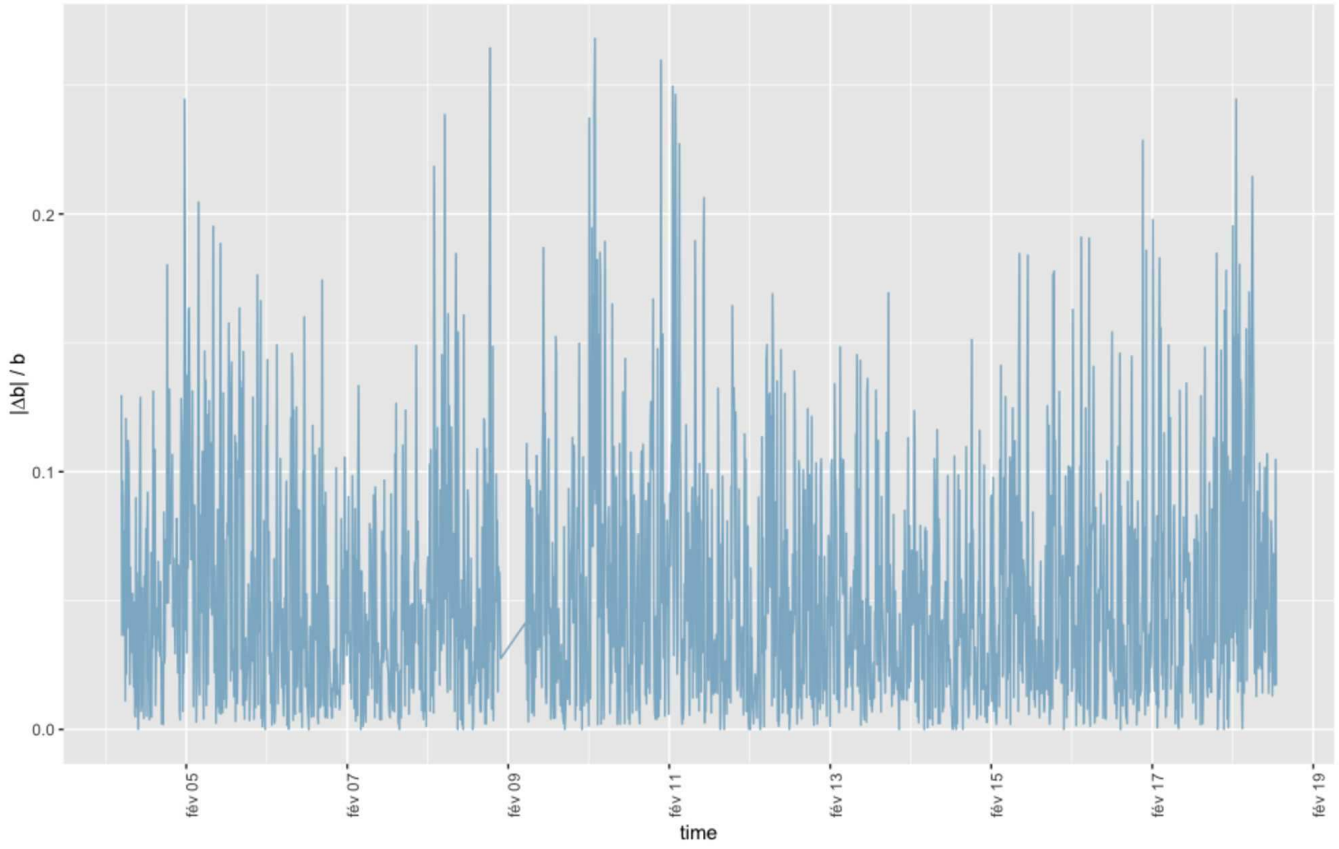

FIGURE 19 : **Stabilité temporelle du maximum de la centralité de chemin.** Le graphe montre dans le temps la dérivée normalisée du maximum de la centralité de chemin, qui capture ses variations relatives à chaque pas de temps. La valeur maximale de 25% correspond à de très fortes perturbations du réseau sur les liens correspondants, puisque cela implique qu'au moins cette proportion d'utilisateurs prenant le lien dans des conditions précédentes doivent prendre un trajet complètement différent.

HÉTÉROGÉNÉITÉ SPATIALE DE L'ÉQUILIBRE Afin d'obtenir un point de vue différent sur la variabilité spatiale des motifs de congestion, nous proposons d'utiliser un indice d'auto-corrélation spatiale, l'indice de Moran (défini par exemple dans [TSAI, 2005]). Utilisé plus généralement en analyse spatiale, avec diverses applications allant de l'étude de la forme urbaine à la quantification de la ségrégation, il peut être appliqué à toute variable spatiale. Il permet d'établir des relations de voisinage et révèle la consistence spatiale locale d'un équilibre s'il est appliqué à une variable de trafic localisée. À un point donné de l'espace, l'auto-corrélation locale pour la variable c est calculée par

$$\rho_i = \frac{1}{K} \cdot \sum_{i \neq j} w_{ij} \cdot (c_i - \bar{c})(c_j - \bar{c}) \tag{4}$$

où K est une constante de normalisation égale à la somme des poids spatiaux fois la variance de la variable et $\bar{c}$ est la moyenne de la variable. Dans notre cas, nous choisissons des poids spatiaux de la forme $w_{ij} = \exp\left(\frac{-d_{ij}}{d_0}\right)$ avec d_0 distance typique de décroissance. L'auto-corrélation est calculée sur la congestion des liens, localisée au centre du lien. Elle capture ainsi les corrélations spatiales dans un rayon du même ordre que la distance de décroissance autour du point i. La moyenne sur l'ensemble des points fournit l'indice d'auto-corrélation spatiale I. Une stationnarité des flots devrait impliquer une stabilité temporelle de l'index.

La figure 20 présente l'évolution temporelle de l'auto-corrélation spatiale pour la congestion. Comme attendu, on observe une forte décroissance de l'auto-corrélation avec la distance de décroissance, à la fois sur l'amplitude et les moyennes temporelles. La forte variabilité temporelle implique de courtes échelles temporelles pour des fenêtres potentielles de stationnarité. Pour une distance de décroissance de 1km, en comparant l'auto-corrélation à la congestion (ajustée à l'échelle du graphe pour lisibilité), on observe que les fortes corrélations coincident avec les heures creuses, tandis que les heures de pointe correspondent à une décroissance des corrélations.

Notre interprétation, combinée avec la variabilité observée des motifs spatiaux, est que les heures de pointe correspondent à un comportement chaotique du système, puisque les bouchons peuvent émerger dans n'importe quel lien du réseau : la corrélation disparait alors puisque l'espace des phases atteignables pour un système dynamique chaotique est rempli uniformément par les trajectoires, de façon équivalente à des vitesses relatives qui apparaitraient comme aléatoires et indépendantes.

Nous avons décrit une étude empirique ayant pour but une approche simple, mais selon notre point de vue nécéssaire, de l'existence de l'équilibre utilisateur statique, plus précisément de sa stationnarité

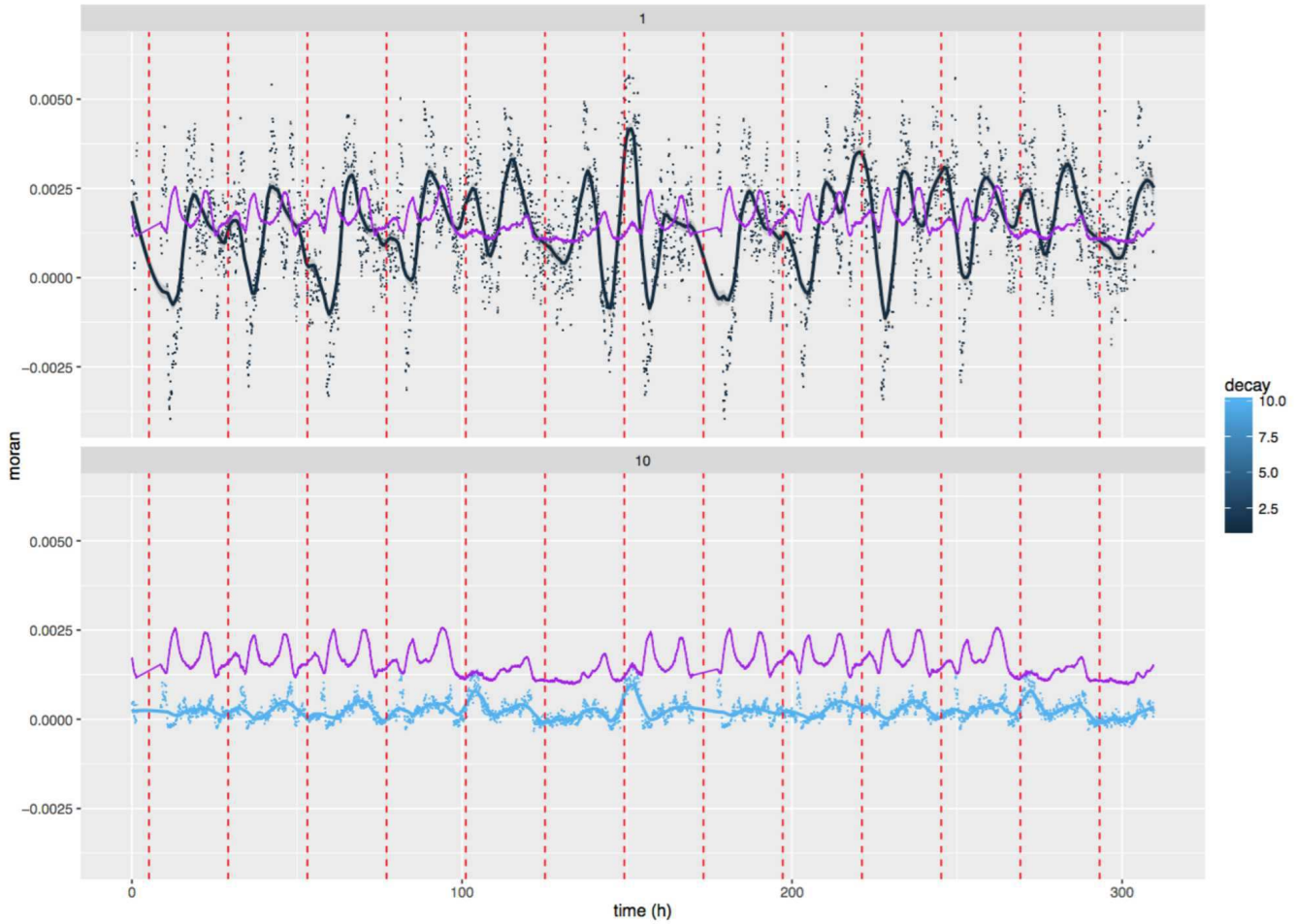

FIGURE 20 : **Auto-corrélations spatiales pour les vitesses relatives sur deux semaines.** Le graphe montre les valeurs de l'auto-corrélation dans le temps, pour des valeurs variables (1km ,10km) de la distance de décroissance (données en couleur et titre des graphes). les valeurs intermédiaires de la distance de décroissance donnent une déformation relativement continue entre ces deux extrêmes. Les points sont lissés sur une fenêtre temporelle de 2h pour faciliter la lecture. Les lignes pointillées verticales correspondent à minuit de chaque jour. La courbe violette donne la vitesse relative, ajustée à l'échelle pour établir la correspondance entre les heures de pointe et les variations de l'auto-corrélation.

dans le temps et l'espace pour un réseau routier métropolitain principal. Un jeu de données de congestion du trafic est construite par collection de données, pour la métropole parisienne sur 3 mois avec une granularité temporelle de 2 minutes. L'exploration interactive du jeu de données via une application web permettant la visualisation spatio-temporelle aide à guider les analyses quantitatives. La variabilité spatio-temporelle des plus courts chemins et de la topologie du réseau, en particulier la centralité de chemin, révèle que l'hypothèse de stationnarité ne tient généralement pas, ce qui est confirmé par l'étude de l'auto-corrélation spatiale de la congestion du réseau.

Mise en perspective

Nous pouvons proposer une mise en perspective de ce travail au regard de notre problématique générale de la co-évolution. Les flux de trafic, qui sont représentatifs du fonctionnement du réseau de transport, sont générés par la distribution spatiale des activités et les comportements des agents microscopiques. Or nous venons de montrer que l'évolution temporelle de ces flux est complexe, rappelant des dynamiques chaotiques, ce qui peut être également compris comme un rôle essentiel de la non-linéarité dans l'émergence de la congestion.

Comme nous l'avons montré au chapitre 1, ces processus liés à la mobilité quotidienne sont probablement liés à un niveau propre de co-évolution entre réseaux et territoires (par exemple la congestion induisant une évolution du réseau, mais aussi éventuellement des relocalisations), que nous n'abordons pas en profondeur dans notre travail. Cette illustration montre ainsi (i) une illustration des interactions entre réseaux et territoires à cette échelle microscopique, suggérant l'existence d'effets complexes à cette échelle ; (ii) la difficulté éventuelle d'une modélisation de la co-évolution à cette échelle microscopique, vu les trajectoires chaotiques du système étudié.

Cette exploration empirique nous a permis d'illustrer d'une part la construction d'un jeu de données ouvertes pour combler l'absence de données, et d'autre part le rôle crucial de l'exploration interactive, qui doit rester combinée à des analyses plus poussées guidées par celle-ci.

⋆ ⋆

⋆

Nous avons ainsi détaillé dans cette section certains enjeux liés à la reproductibilité et à la science ouverte, complétant notre positionnement spécifiques en termes de modélisation à un positionnement plus général correspondant à la pratique scientifique.

Nous allons finalement dans la dernière section qui suit encore monter en généralité et préciser nos positionnement épistémologiques, c'est-à-dire concernant les disciplines elles-mêmes et la production de connaissance. Cette étape sera cruciale, puisque notre positionnement au regard des systèmes sociaux et des systèmes biologiques permettra d'introduire les éléments fondamentaux pour une définition plus complète de la co-évolution.

⋆ ⋆

⋆

3.3 POSITIONNEMENT ÉPISTÉMOLOGIQUE

La dernière section de ce chapitre vise à clarifier notre positionnement épistémologique, celui-ci ayant déjà été ébauché à plusieurs occasions précédemment. Un tel positionnement n'est jamais anodin, puisqu'il conditionne fortement les démarches, les expériences et l'interprétation des résultats : comme le souligne [MORIN, 1980], un positionnement qui se dit objectif en rejetant toute subjectivité est bien plus biaisé qu'une approche subjective consciente.

Les points que nous souhaitons développer se placent dans une logique à la fois verticale de niveau d'abstraction et dans une logique de domaines scientifiques : dans l'ordre, nous posons d'abord le contexte épistémologique général (propre à l'histoire des sciences, à un niveau d'abstraction moyen), pour descendre en généralité et préciser conceptuellement nos objets particuliers (épistémologie du vivant et du social), pour finalement tout remettre en perspective au niveau de la production de connaissance elle-même (épistémologie de la complexité).

3.3.1 *Approche cognitive et perspectivisme*

Notre positionnement épistémologique se fonde sur une approche cognitive de la science, introduite par GIERE dans [GIERE, 2010b]. L'approche se concentre sur le rôle des agents cognitifs comme porteurs et producteurs de la connaissance. Son caractère opérationnel a été montré par [GIERE, 2010a] qui étudie un modèle multi-agents de la science. Ces idées convergent avec le jeu Nobel de CHAVALARIAS [CHAVALARIAS, 2016] qui teste de manière stylisée l'équilibre entre production de nouvelles théories et tentative de falsification de théories existantes dans l'entreprise scientifique collective.

Ce positionnement épistémologique a été présenté par GIERE comme *perspectivisme scientifique* [GIERE, 2010c], dont la caractéristique principale est de considérer toute entreprise scientifique comme une *perspective* dans laquelle des *agents* utilisent des *media* (modèles) pour représenter quelque chose dans un certain but. Pour comprendre ses principes de manière plus concrète, nous pouvons le positionner sur la *check-list* du constructivisme de HACKING [HACKING, 1999], un outil pratique pour situer une position épistémologique. Celle-ci suppose un espace simplifié tri-dimensionnel dans lequel les dimensions sont différents aspects sur lesquels les approches réalistes et constructivistes généralement divergent. Le premier point est le niveau de contingence (dépendance au chemin du processus de construction de connaissances) : celle-ci est nécessaire dans l'approche perspectiviste qui est pluraliste et suppose des chemins parallèles de construction de connaissance. Le deuxième point mesure un "degré de constructivisme", qui est assez haut en perspectivisme car les agents produisent

la connaissance. Enfin, le dernier point qui concerne l'explication endogène ou exogène de la stabilité des théories, est fortement du côté du constructivisme, puisque cette stabilité dépend des interactions complexes entre les agents et leur perspectives et donc totalement endogène. Le perspectivisme a pour ces raisons été présenté comme un chemin intermédiaire et alternatif entre le réalisme absolu et le constructivisme sceptique [Brown, 2009]. Le concept de *perspective* jouera pour nous un rôle fondamental dans le cadre développé en 8.3.

Cette approche mettant l'emphase sur l'auto-organisation, nous la voyons totalement compatible avec une vision anarchiste de la science comme défendue par [Feyerabend, 1993]. Celui-ci émet des doutes sur l'intérêt de l'anarchisme politique mais introduit l'*anarchisme scientifique*, qu'il ne faut pas comprendre comme un refus total de toute méthode "objective", mais d'une autorité et légitimité artificielle que certaines méthodes ou courants scientifiques pourraient vouloir prendre. Il démontre par une analyse précise des travaux de Galilée que la plupart de ses résultats étaient basés sur des croyances et que la plupart n'étaient pas accessibles avec les outils et méthodes de l'époque, et postule qu'il devrait en être de même pour certains travaux contemporains. Il n'y a donc pas de *perspective* objectivement plus légitime que d'autres dans la mesure de leurs validation par des faits et des pairs - et même dans ces cas la légitimité doit pouvoir être discutée, car la remise en question est un fondement de la connaissance. Cela correspond exactement à la pluralité des perspectives que nous défendons.

Supposer auto-organisation et émergence des connaissances peut être interprété comme une priorité donnée à la construction des paradigmes *par le bas* (*bottom-up*), en tentant de se distancer des préconceptions ou dogmes cadrant par le haut. En d'autres termes, il s'agit de pratiquer l'anarchisme scientifique prôné par Feyerabend. En effet, les positions anarchistes ont trouvé un écho très cohérent dans les différents courants de la complexité, de la cybernétique à l'auto-organisation au cours du 20ème siècle [Duda, 2013]. Notre cadre de connaissances développé en 8.3 illustre cette émergence de la connaissance. De plus, notre volonté de réflexivité et de donner à notre travail des pistes de lecture diverses au delà de la linéarité (voir Annexe F), montre l'application de ces principes. Les recommandations méthodologiques et les positionnements donnés précédemment dans ce chapitre pourraient sonner comme totalitaires s'ils étaient assénés de manière sèche sans contexte, mais ceux-ci sont en fait tout le contraire puisqu'ils découlent d'une dynamique récente de science ouverte qui a bien émergé par le bas, conséquence en partie de l'ouverture et de la pluralité.

3.3.2 *De la Vie à la Culture*

Systèmes biologiques et systèmes sociaux

Le parallèle entre les systèmes sociaux et les systèmes biologiques est souvent fait, parfois de manière plus qu'imagée comme par exemple pour la théorie du *Scaling* de WEST qui applique des équations de croissance similaires à partir des lois d'échelle, avec des conclusions inverses tout de même concernant la relation entre taille et rythme de vie [BETTENCOURT et al., 2007]. Les relations d'échelle ne tiennent plus lorsqu'on essaye de les appliquer à une fourmi seule, et il faut alors l'appliquer à la fourmilière entière qui est l'organisme en question. En ajoutant la propriété de cognition, on confirme qu'il s'agit du niveau pertinent, puisque celle-ci possède des propriétés cognitives avancées, comme la résolution de problèmes d'optimisation spatiaux, ou la réponse rapide à une perturbation extérieure. Les organisations sociales humaines, les villes, peuvent-elles être vues comme des organismes ? [BANOS, 2013] file la métaphore de la *fourmilière urbaine* mais rappelle que le parallèle s'arrête assez vite. Nous allons voir cependant dans quelle mesure certains concepts de l'épistémologie de la biologie peuvent être utiles pour comprendre les systèmes sociaux que nous nous proposons d'étudier.

Nous nous basons sur la contribution fondamentale de MONOD dans [MONOD, 1970], qui tente de développer les principes épistémologiques cruciaux pour l'étude du vivant. Ainsi, les organismes vivants répondent à trois propriétés essentielles qui permettent des les différencier d'autres systèmes : (i) la téléonomie, c'est-à-dire qu'il s'agit "d'objets doués d'un projet", projet qui se reflète dans leur structure et dans celles des artefacts qu'ils produisent[39] ; (ii) l'importance des processus morphogénétiques dans leur constitution (voir 5.1) ; (iii) la propriété de reproduction invariante de l'information définissant leur structure. MONOD esquisse de plus en conclusion des pistes pour une théorie de l'évolution culturelle. La téléonomie est essentielle dans les structures sociales, puisque toute organisation essaye de satisfaire un ensemble d'objectifs, même si en général elle n'y parviendra pas et que ceux-ci co-évolueront avec l'organisation. Cette notion d'optimisation multi-objectif est typique des systèmes complexes socio-techniques, et y sera plus cruciale que pour les systèmes biologiques.

Ensuite, nous postulons que le concept de morphogenèse est un outil essentiel pour comprendre ces systèmes, avec une définition très proche de celle utilisée en biologie. Un travail approfondi pour donner cette définition est fait en 5.1, que nous résumerons en l'existence de processus relativement autonomes guidant la croissance du système et impliquant des relations causales circulaires entre forme

39 Qu'il ne faut pas confondre avec la téléologie, propres aux animismes, qui consiste à prêter un projet ou un sens à l'univers.

et fonction qui témoignent d'une architecture émergente. Pour des systèmes sociaux, isoler le système est plus difficile et la notion de frontière sera moins stricte que pour un système biologique, mais on retrouvera bien ce lien entre forme et fonction, comme par exemple la structure d'une organisation ayant un impact sur ses fonctionnalités.

Enfin, la reproduction de l'information est au coeur de l'évolution culturelle, par la transmission de la culture et la *mémétique*, la différence étant que le rapport d'échelles de temps entre la fréquence de transmission et les processus de croisement et de mutation ou d'autres processus non mémétiques de production culturelle est relativement faible, alors qu'elle est de plusieurs ordres de grandeur en biologie.

Un exemple illustre que le parallèle n'est pas toujours absurde : [Gabora et Steel, 2017] propose un modèle de réseau auto-catalytique pour la cognition, qui expliquerait l'apparition de l'évolution culturelle par des processus analogues à ceux s'étant produit à l'apparition de la vie, c'est-à-dire une transition permettant au molécules de s'auto-entretenir et s'auto-reproduire, les représentations mentales faisant office de molécules.

Mais si les processus à l'origine sont analogues, la nature de l'évolution est bien différente par la suite, comme le montrent [Leeuw, Lane et Read, 2009], les critères darwiniens d'évolution n'étant pas suffisant pour expliquer l'évolution de nos sociétés organisées. Il s'agit d'une complexité de nature différente dans laquelle le rôle des flux d'information est crucial (voir le rôle de la complexité informationnelle dans la sous-section suivante).

L'un des points sur lequel il s'agit également d'être attentif est la plus grande difficulté de définir les niveaux d'émergence pour les systèmes sociaux : [Roth, 2009] souligne le risque de tomber dans des cul-de-sac ontologiques car les niveaux ont été mal définis. Il soutient qu'il faut d'une manière générale penser au-delà de la seule dichotomie micro-macro qui est utilisée pour caricaturer les concepts d'émergence faible, et que les ontologies doivent souvent être multi-niveaux et impliquant de multiples niveaux intermédiaires.

Cette dernière question est aussi à mettre en perspective avec le problème de l'existence d'émergence forte dans les structures sociales, qui en termes sociologiques correspond à l'idée de l'existence "d'êtres collectifs" [Angeletti et Berlan, 2015]. Morin distingue d'ailleurs les systèmes vivants du second type (multi-cellulaire) et du troisième type (structures sociales), mais précise que les *sujets* de ces derniers sont nécessairement inachevés [Morin, 1980] (p. 852). Ainsi, les émergences du biologique au social sont analogues mais restent fondamentalement différentes.

Co-évolution

Ce positionnement sur les systèmes biologiques et sociaux trouve un écho immédiat pour le concept de co-évolution. Il provient en effet de la biologie, où il a été développé à la suite de celui d'évolution, pour être utilisé plus récemment en sciences humaines et sociales. Dans quelle mesure le concept a-t-il été transféré ? Retrouve-t-on un parallèle similaire à celui entre évolution biologique et évolution culturelle ? Nous proposons pour répondre à ces questions d'apporter un bref point de vue multidisciplinaire sur la co-évolution[40]. Nous passons par la suite en revue un large spectre de disciplines, partant de la biologie où le concept a initialement trouvé son origine pour arriver progressivement à des disciplines en relation avec les sciences du territoire.

Biologie

Le concept de co-évolution en biologie est une extension de celui bien connu d'*évolution*, qui remonte à DARWIN. [DURHAM, 1991] (p. 22) rappelle les composantes et structure systémiques nécessaires pour qu'il y ait évolution[41].

1. Processus de *transmission*, impliquant des unités de transmission et des mécanismes de transmission.
2. Processus de *transformation*, nécessitant des sources de variation.
3. Isolation de sous-systèmes pour que les effets des processus précédents soient observables dans des différentiations.

Ainsi, une population soumise à des contraintes (souvent synthétisées conceptuellement comme une *fitness*) qui conditionnent la transmission du patrimoine génétique des individus (transmission), et à des mutation génétiques aléatoires (transformation), sera bien en évolution dans les territoires spatiaux qu'elle occupe (isolation), et par extension l'espèce à laquelle on peut l'associer.

40 La démarche ici est légèrement différente de celle que nous mènerons en 5.1 dans le cas de la morphogenèse, qui sera *interdisciplinaire* au sens où elle cherchera à intégrer les approches, tandis que nous restons ici dans un aperçu des concepts et donc plutôt dans du *multidisciplinaire*. Le concept de *co-évolution* étant clé pour notre travail empirique par la suite, nous en donnerons alors une caractérisation originale et prenons le parti de ne pas tomber dans le syncrétisme intégrateur pour ce concept, mais bien de l'approcher d'un *point de vue géographique*, et même plus précisément dans le cadre des systèmes territoriaux. On pourrait postuler une congruence entre la spécialisation empirique/de modélisation et celle théorique, plaçant notre processus de production de connaissance dans un profil particulier de dynamiques de domaines de connaissance (voir 8.3).

41 Et dans ce contexte général l'évolution n'est pas réservée à la biologie du vivant et la présence de gènes, mais aussi à des systèmes physiques vérifiant ces conditions. Nous y reviendrons plus loin.

La co-évolution est alors définie comme un changement évolutionnaire dans une caractéristique des individus d'une population, en réponse à un changement dans une deuxième population qui à son tour répond évolutionnairement au changement de la première, comme synthétisé par [Janzen, 1980]. Cet auteur appuie par ailleurs la subtilité du concept et alerte contre ses utilisations injustifiées : la présence d'une congruence de deux caractéristiques qui semblent adaptées l'une à l'autre n'implique pas l'existence d'une co-évolution, l'une des deux espèces ayant pu s'adapter seule à une caractéristique déjà présente de l'autre.

Cette présentation brute de décoffrage mutile dans une certaine mesure la complexité réelle des écosystèmes : les populations s'insèrent dans des réseaux trophiques et des environnements, et les interactions co-évolutionnaires impliqueraient des communautés de populations d'espèces diverses, comme présenté par [Strauss, Sahli et Conner, 2005] sous l'appellation de co-évolution diffuse. De même, les dynamiques spatio-temporelles sont cruciales dans la réalisation de ces processus : [Dybdahl et Lively, 1996] étudient par exemple l'influence de la distribution spatiale sur les motifs de co-évolution pour un escargot et son parasite, et montrent qu'une vitesse de diffusion génétique dans l'espace plus grande pour le parasite conduit les dynamiques de co-évolution.

Les concepts essentiels à retenir du point de vue biologique sont ainsi : (i) existence de processus d'évolution, en particulier transmission et transformation ; (ii) dans des schémas circulaires entre populations dans le cas de la co-évolution ; et (iii) dans un cadre territorial (spatio-temporel et environnemental au sens du reste de l'écosystème) complexe.

Evolution culturelle

Ce développement sur la co-évolution nous a été amené par le parallèle entre systèmes biologiques et systèmes sociaux. L'évolution de la culture est théorisée est explorée par un champ propre, et n'est pas en reste de dynamiques co-évolutives. [Mesoudi, 2017] rappelle l'état des connaissances sur le sujet et les défis à venir, comme la relation avec la nature cumulative de la culture, l'influence de la démographie dans les processus d'évolution, ou la construction de méthodes phylogénétiques permettant de reconstruire des arbres des branchements passés.

Pour donner un exemple, [Carrignon, Montanier et Rubio-Campillo, 2015] introduit un cadre conceptuel pour la co-évolution de la culture et du commerce dans le cas de sociétés anciennes sur lesquelles on dispose de données archéologiques, et propose son implémentation par un modèle multi-agents dont les dynamiques sont partiellement validées par l'étude des faits stylisés produits par le modèle. La co-évolution est bien prise ici au sens d'adaptation mutuelle de struc-

tures socio-spatiales, à des échelles de temps comparables, dans ce cadre plus général d'évolution culturelle.

L'évolution culturelle serait même indissociable de l'évolution génétique, puisque [Durham, 1991] postule et illustre un lien fort entre les deux, qui seraient eux-mêmes en co-évolution. [Bull, Holland et Blackmore, 2000] explore un modèle stylisé impliquant deux populations de répliquants (les gènes et les memes) et montre l'existence de transitions de phase pour les résultats du processus d'évolution génétique lorsque l'interaction avec le répliquant culturel est forte.

Sociologie

Le concept a été utilisé en sociologie et disciplines apparentées comme les études de l'organisation, suivant le parallèle effectué ci-dessus de la même manière que pour l'évolution culturelle. Dans le domaine de l'étude des organisations, [Volberda et Lewin, 2003] développent un cadre conceptuel de la co-évolution inter-organisationnelle en relations avec les processus de management internes, mais déplore l'absence d'études empiriques cherchant à quantifier cette co-évolution. Dans le cadre de la gestion des systèmes de production, [Tolio et al., 2010] conceptualisent un chaine de production intelligente où produit, processus et système de production doivent être en co-évolution.

Economie géographique

En économie géographique, le concept de co-évolution a également largement été mobilisée. L'idée d'entités évolutionnaires en économie vient à contre-courant du courant néoclassique qui reste majoritaire, mais trouve un écho de plus en plus pertinent [Nelson et Winter, 2009]. [Schamp, 2010] procède à une analyse épistémologique de l'utilisation de la co-évolution, et oppose une approche néo-schumpeterienne de l'économie qui considère l'émergence de populations qui évoluent à partir de règles micro-économiques (qui correspondrait à une lecture directe et relativement isolationniste de l'évolution biologique) à une approche systémique qui considérerait l'économie comme un système évolutif de manière globale (qui correspondrait à l'évolution diffuse que nous avons développé précédemment), pour proposer une caractérisation précise tombant dans le premier cas, qui suppose des *institutions* qui co-évoluent. Le plus important pour notre propos est qu'il souligne l'aspect crucial du choix des population et des entités considérées, de la zone géographique, et appuie l'importance de l'existence de relations causales circulaires.

Il est possible de donner divers exemples d'application. [Wal et Boschma, 2011] introduisent un cadre conceptuel pour permettre de concilier nature évolutionnaire des entreprises, théorie des clusters et réseaux de connaissance, dans lequel la co-évolution entre réseaux et

entreprises est centrale, et qui est définie comme une causalité circulaire entre différentes caractéristiques de ces sous-systèmes. [Colletis, 2010] introduit un cadre de co-évolution des territoires et de la technologie (questionnant par exemple le rôle de la proximité pour les innovations), qui révèle l'importance à nouveau de l'aspect institutionnel. Le cadre proposé par [Ter Wal et Boschma, 2011] couple la vision évolutionnaire des entreprises, la littérature sur les industries et l'innovation dans les clusters, et l'approche par réseau complexe des connexions entre ces premiers dans le système territorial.

En économie environnementale, [Kallis, 2007] montre que des approches "larges" (pouvant considérer la majorité des co-dynamiques comme co-évolutives) s'opposent à des approches plus strictes (dans l'esprit de la définition donnée par [Schamp, 2010]), et que dans tous les cas une définition précise, ne venant pas forcément de la biologie, doit être donnée, en particulier pour la recherche d'une caractérisation empirique.

Géographie

Pour la géographie, comme nous l'avons déjà présenté en introduction, les travaux les plus proches empiriquement et théoriquement des notions de co-évolution sont étroitement liés à la théorie évolutive des villes. Il n'est pas évident de tracer dans la littérature à quel moment la notion a été clairement formalisée, mais il est évident qu'elle était présente dès les fondements de la théorie comme le rappelle Denise Pumain (voir D.3) : le système complexe adaptatif est composé de sous-systèmes en interdépendances complexes, souvent circulairement causales. Les premiers modèles incluent bien cette vision de manière implicite, mais la co-évolution n'est pas appuyée explicitement ou définie précisément, en termes qui seraient quantifiables ou identifiables structurellement. [Paulus, 2004] amène des preuves empiriques de mécanismes de co-évolution par l'étude de l'évolution des profils économiques des villes françaises. L'interprétation utilisée par [Schmitt, 2014] repose sur une entrée par la théorie évolutive des villes, et consiste fondamentalement en une lecture des systèmes de villes comme entités fortement interdépendantes.

Géographie physique

En étude des paysages, [Sheeren et al., 2015] parlent de co-évolution du paysage et des activités agricoles, mais ne considèrent en fait pas d'effet circulaires de l'un sur l'autre. A priori, leurs résultats montrent que l'évolution des pratiques agricoles entraine une évolution du paysage, et il n'est ainsi pas clair dans quelle mesure le cadre conceptuel de la co-évolution, mentionné sans plus de détails, est mobilisé.

Physique

Enfin, on peut noter de manière anecdotique que le terme de co-évolution a également été utilisé par la physique. L'utilisation pour des systèmes physiques peut porter à débat, selon que l'on suppose ou non que la transmission suppose une transmission d'*information*[42]. Dans le cas d'une transmission ontologique uniquement physique (*êtres physiques*), alors une grande partie des systèmes physiques sont évolutifs. [HOPKINS et al., 2008] développent un cadre cosmologique pour la co-évolution d'objets cosmiques hétérogènes dont la présence et les dynamiques sont difficilement expliquées par des théories plus classiques (certains types de galaxies, quasars, trous noirs supermassifs). [ANTONIONI et CARDILLO, 2017] étudient la co-évolution entre des propriétés de synchronisation et de coopération au sein d'un réseau d'oscillateurs de Kuramoto[43], montrant d'une part que le concept peut être appliqué à des objets abstraits, et d'autre part qu'un réseau de relations complexes entre variables peut être à l'origine de dynamiques présentant des causalités circulaires, c'est-à-dire d'une co-évolution en ce sens.

Synthèse

La plupart de ces approches rentrent dans la théorie des systèmes complexes adaptatifs développée par HOLLAND, notamment dans [HOLLAND, 2012] : il voit tout système comme une imbrication de systèmes de limites, filtrant des signaux ou des objets. Au sein d'une limite donnée, le sous-système correspondant est relativement autonome de l'extérieur, est est appelé *niche écologique*, en correspondance directe avec les communautés fortement connectées au sein des réseaux trophiques ou écologiques. Ainsi, des entités interdépendantes au sein d'une niche sont dites en co-évolution. Nous reviendrons sur cette entrée lors de la construction théorique en 8.2 lorsque nous aurons développé d'autres concepts qui lui sont nécessaire.

Nous retenons de cet aperçu multidisciplinaire de la co-évolution les points fondamentaux suivants précurseurs à une définition propre de la co-évolution que nous donnerons plus loin, en conclusion de la première partie.

---

42 L'information est définie dans la théorie shanonienne comme une probabilité d'occurrence d'une chaîne de caractère. [MORIN, 1976] montre que le concept d'information est en fait bien plus complexe, et qu'il doit être pensé conjointement à un contexte donné de génération d'un système auto-organisateur néguentropique, i.e. réalisant des diminutions locales d'entropie notamment grâce à cette information. Ce type de système est nécessairement vivant. Nous prendrons ici cette vision complexe de l'information.

43 Le modèle de Kuramoto s'intéresse à la synchronisation au sein de systèmes complexes, en étudiant l'évolution de phases θ_i couplée par les équations d'interaction $\ddot{\vec{\theta}} = \vec{\omega} + \vec{W}\left[\vec{\theta}\right] + \mathbf{B}$ où $\vec{\omega}$ sont les phases propres de forçage et la force de couplage entre i et j est donnée par $\vec{W}_i = \sum_j w_{ij} \sin\left(\theta_i - \theta_j\right)$ et $\vec{B}$ du bruit.

1. La présence de *processus d'évolution* est primaire, et leur définition se ramène presque toujours à l'existence de processus de transmission et de transformation.

2. La co-évolution suppose des entités ou systèmes, appartenant à des classes distinctes, dont les dynamiques évolutives sont couplées de manière circulaire causale. Les approches peuvent différer selon l'hypothèse de populations de ces entités, d'objets singuliers, ou de composantes d'un système global alors en interdépendance mutuelle sans qu'il y ait circularité directe.

3. La délimitation des systèmes ou des sous-systèmes, à la fois dans l'espace ontologique (définition des objets étudiés), mais aussi dans l'espace et le temps, ainsi que leur distribution dans ces espaces, est fondamental pour l'existence de dynamiques co-évolutives, et a priori dans un grand nombre de cas, pour leur caractérisation empirique.

3.3.3 *Nature de la complexité et production de connaissances*

Les deux premiers points épistémologiques que nous venons de traiter relevaient respectivement du positionnement en lui-même, c'est-à-dire du cadre de lecture des processus de production de connaissance scientifique, puis de la nature des concepts considérés. Nous proposons de monter encore en généralité par rapport au premier et d'introduire un développement contribuant modestement (c'est-à-dire dans notre contexte) à *la connaissance de la connaissance*. Il s'agit d'interroger les liens entre complexité et processus de production de connaissances.

Un aspect de la production de connaissance sur des systèmes complexes, auquel nous nous heurtons plusieurs fois ici (voir chapitre 8), et qui semble être récurrent voire inévitable, est un certain niveau de réflexivité (et qui serait inhérent aux systèmes complexes en comparaison aux systèmes simples, comme nous le développerons plus loin). Nous entendons par là à la fois une réflexivité pratique, c'est-à-dire la nécessité d'élever le niveau d'abstraction, comme le besoin de reconstruire de manière endogène les disciplines dans lesquelles une réflexion cherche à se positionner comme proposé en 2.2, ou de réfléchir à la nature épistémologique de la modélisation lors de l'élaboration d'un modèle comme en B.5, mais également une réflexivité théorique en le sens que les appareils théoriques ou les concepts produits peuvent s'appliquer de manière récursive à eux-mêmes. Cette constatation pratique fait écho à des débats épistémologiques anciens questionnant la possibilité d'une connaissance objective de l'univers qui serait indépendante de notre structure cognitive, ou bien la nécessité d'une "rationalité évolutive" impliquant que notre système cognitif, produit de l'évolution, reflète les processus complexes ayant

conduit à son émergence, et que toute structure de connaissance sera par conséquent réflexive[44]. Nous ne prétendons pas ici apporter une réponse à une question aussi vaste et vague telle quelle, mais proposons un lien potentiel entre cette réflexivité et la nature de la complexité.

Complexité et complexités

Ce qui est entendu par complexité d'un système mène souvent à des malentendus car celle-ci peut être qualifiée selon différentes dimensions et visions. Nous distinguons dans un premier temps la complexité au sens d'émergence faible et d'autonomie entre les différents niveaux d'un système, et sur laquelle différentes positions peuvent être développées comme dans [Deffuant et al., 2015]. Nous ne rentrerons pas dans une granularité plus fine, la vision de la complexité sociale donnant encore plus de fil à retordre au démon de Laplace, et pouvant être par exemple comprise par une émergence plus forte (au sens d'émergence faible et forte développée précédemment en 3.1). Nous simplifions ainsi et supposons que la nature des systèmes joue un rôle secondaire dans notre reflexion, et considérons la complexité au sens d'une émergence.

D'autre part, nous distinguons deux autres "types" de complexité, la complexité computationnelle et la complexité informationnelle, qui peuvent être vues comme des mesures de complexité, mais qui ne sont pas directement équivalentes à l'émergence, puisqu'il n'existe pas de lien systématique entre les trois. On peut par exemple imaginer utiliser un modèle de simulation, pour lequel les interactions entre agents élémentaires se traduisent par un message codé au niveau supérieur : il est alors possible en exploitant les degrés de liberté de minimiser la quantité d'information contenue dans le message. Les différentes langues demandent des efforts cognitifs différents et compressent différemment l'information, ayant différents niveau de complexité mesurables [Febres, Jaffé et Gershenson, 2013]. De même, des artefacts architecturaux sont le résultat d'un processus d'évolution naturelle puis culturelle et peuvent témoigner plus ou moins de cette trajectoire.

De nombreuses autres caractérisations conceptuelles ou opérationnelles de la complexité existent, et il est clair que la communauté scientifique n'a pas convergé sur une définition unique [Chu, 2008][45]. Nous proposons de nous concentrer sur ces trois concepts en particulier, pour lesquels les relations ne sont déjà pas évidentes.

44 Nous remercions D. Pumain d'avoir pointé cette vue alternative du problème que nous allons développer par la suite.

45 Dans une approche en un sens réflexive, [Chu, 2008] propose de continuer d'explorer les différentes approches existantes, comme des proxys de la complexité dans le cas d'un essentialisme, ou comme des concepts à part entière. La complexité devrait émerger d'elle même de l'interaction entre ces différentes approches étudiant la complexité, d'où la réflexivité.

En effet, les liens entre ces trois types de complexité ne sont pas systématiques, et dépendent du type de système. Des liens épistémologiques peuvent néanmoins être introduits. Nous traitons ceux entre émergence et les deux autres complexités, étant donné que le lien entre complexité computationnelle et complexité informationnelle est assez bien compris et relève de problématiques de compression de l'information et de traitement du signal, ou encore de cryptographie.

Complexité computationnelle et émergence

Différents indices suggèrent une certaine nécessité de complexité computationnelle pour avoir émergence dans des systèmes complexes, tandis que réciproquement un certain nombre de systèmes complexes adaptatifs sont dotés de capacités de calcul élevées.

Un premier lien où complexité computationnelle implique émergence est suggéré par un examen algorithmique des problèmes fondamentaux de la physique quantique. En effet, [Bolotin, 2014] démontre que la résolution de l'équation de Schrödinger avec Hamiltonien quelconque est un problème NP-difficile et NP-complet, et donc que l'acceptation de $\mathbf{P} \neq \mathbf{NP}$ implique une séparation qualitative entre le niveau quantique microscopique et le niveau d'observation macroscopique. Ainsi, c'est bien la complexité (ici au sens de leur calcul) des interactions au sein du système et de son environnement qui explique l'apparente réduction du paquet d'onde, ce qui rejoint l'approche de Gell-Mann par la décohérence quantique [Gell-Mann et Hartle, 1996], qui explique que des probabilités ne peuvent être associées qu'aux histoires décohérentes (dans lesquelles les corrélations ont fait prendre une trajectoire au système à l'échelle macroscopique)[46]. Le paradoxe du chat de Schrödinger nous apparait ainsi comme une perspective fondamentalement réductionniste, puisqu'il suppose que la superposition d'états peut se propager à travers les niveaux successifs et qu'il n'y aurait pas émergence, au sens de constitution d'un niveau supérieur autonome. En d'autres termes, le travail

46 Le *Problème de la Mesure Quantique* se pose lorsqu'on considère une fonction d'onde microscopique donnant l'état d'un système pouvant être superposition de plusieurs états, et consiste en un paradoxe théorique, les mesures étant toujours déterministes alors que le système a des probabilité d'états d'une part, et le problème de la non-existence d'états macroscopiques superposés (réduction du paquet d'onde) d'autre part. Comme revu par [Schlosshauer, 2005], différentes interprétations épistémologiques de la physique quantiques sont liées à différentes explications de ce paradoxe, dont celle "classique" de Copenhague qui donne à l'acte d'observation le rôle de reduction du paquet d'onde. Gell-Mann précise que cette interprétation n'est pas absurde puisque c'est bien les corrélations entre l'objet quantique et le monde qui produisent l'histoire décohérente, mais qu'elle est bien trop spécifique, et que la réduction a lieu dans l'émergence elle-même : le chat est bien mort ou vivant, mais pas les deux, avant que l'on ouvre la boîte.

de [Bolotin, 2014] suggère que la complexité computationnelle est suffisante pour la présence d'émergence[47].

Dans le sens inverse, le lien entre complexité computationnelle et émergence est mis en valeur par les questions liées à la nature de la computation [Moore et Mertens, 2011]. Des automates cellulaires, qui sont par ailleurs cruciaux pour la compréhension de divers systèmes complexes, ont été montrés Turing-complets[48], comme le Jeu de la Vie [Beer, 2004][49]. Des organismes sans système nerveux central sont capables de résoudre des problèmes décisionnels difficiles [Reid et al., 2016]. Un algorithme à base de fourmis est montré par [Pintea, Pop et Chira, 2017] comme résolvant un Problème du Voyageur de Commerce Généralisé (GTSP), problème NP-difficile. Ce lien fondamental avait déjà été envisagé par Turing, puisqu'au delà de ses contributions fondamentales à l'informatique moderne, il s'était intéressé à la morphogenèse et a tenté de produire des modèles chimiques d'explication de celle-ci [Turing, 1952] (qui étaient très loin de effectivement l'expliquer - elle n'est toujours pas bien comprise aujourd'hui, voir 5.1 - mais dont les contributions conceptuelles ont été fondamentales, notamment pour la notion de réaction-diffusion). On sait par ailleurs qu'un minimum de complexité en termes d'interactions constituantes dans un cas particulier de système basé sur les agents (modèles de réseaux booléens), et donc d'émergences possibles, implique une borne inférieure sur la complexité computationnelle, qui devient conséquente dès que les interactions avec l'environnement sont ajoutées [Tošić et Ordonez, 2017].

Complexité informationnelle et émergence

La complexité informationnelle, ou la quantité d'information contenue dans un système et la manière dont celle-ci est stockée, entretient également des liens fondamentaux avec l'émergence. L'information est équivalente à l'entropie d'un système et donc à son degré d'organisation - c'est ce qui a permis de résoudre le paradoxe apparent du

47 A priori, cette séparation effective des échelles n'implique pas que le niveau inférieur ne joue pas un rôle crucial, puisque [Vattay et al., 2015] prouve que les propriétés de criticalité quantiques sont typiques des molécules du vivant, sans qu'il n'y ait a priori de spécificité pour la vie dans cette détermination complexe par les échelles inférieures : [Verlinde, 2017] a introduit une nouvelle approche liant théories quantiques et relativité générale dans laquelle il est montré que la gravité est un phénomène émergent et que la dépendance au chemin dans la déformation de l'espace de base introduit un terme supplémentaire au niveau macroscopique, qui permet d'expliquer les déviations attribuées jusqu'alors à la *matière noire*.

48 Un système est Turing-complet s'il est capable de calculer les mêmes fonctions qu'une machine de Turing, communément accepté comme l'ensemble du "calculable" (thèse de Church). Pour mémoire, une machine de Turing est un automate fini à bande d'écriture infinie [Moore et Mertens, 2011].

49 Il existe même un langage de programmation permettant de programmer en *Game of Life*, disponible à `https://github.com/QuestForTetris`. Sa genèse trouve son origine dans un défi posté sur *codegolf* ayant pour but la conception d'un Tetris, et a abouti à un projet collaboratif extrêmement avancé.

Démon de Maxwell qui serait capable de diminuer l'entropie d'un système isolé et donc contredire la deuxième loi de la thermodynamique : celui-ci utilise en fait l'information sur les positions et vitesses des molécules du système, et son action compense la perte d'entropie par sa captation d'information[50].

Cette notion d'accroissement local de l'entropie a été étudiée largement par Chua sous la forme du *Local Activity Principle*, qui est introduit comme un troisième principe de la thermodynamique, permettant d'expliquer par des arguments mathématiques l'auto-organisation pour une certaine classe de systèmes complexes typiquement impliquant des équations de réaction-diffusion [Mainzer et Chua, 2013].

La manière dont l'information est stockée et compressée est essentielle pour la vie, puisque l'ADN est bien un système de stockage d'information, dont le rôle à différents niveaux est bien loin d'être compris complètement. La complexité culturelle témoigne également d'un stockage de l'information à différents niveaux, par exemple au sein des individus mais aussi des artefacts et des institutions, et des flux d'information relevant nécessairement des deux autres types de complexité. Les flux d'information sont essentiels pour l'auto-organisation dans un système multi-agents. Les comportements collectifs de poissons ou d'oiseaux sont des exemples typiques utilisés pour illustrer l'émergence et font partie des cas d'école de systèmes complexes. On commence cependant seulement à comprendre comment ces flux structurent le système, et quels sont les motifs spatiaux de transfert d'information au sein d'un *flock* par exemple : [Crosato et al., 2017] introduisent des premiers résultats empiriques avec l'entropie de transfert pour des poissons et posent les bases méthodologiques de ce type d'étude.

Production de connaissances

Nous avons à présent la matière suffisante pour en venir à la réflexivité. Il est possible de positionner la production de connaissances à l'intersection des interactions entre types de complexité développées ci-dessus. Tout d'abord, la connaissance telle que nous l'envisageons ne peut se passer d'une construction collective, et implique donc un encodage et une transmission de l'information : il s'agit à un autre niveau de toutes les problématiques liées à la communication scientifique. La production de connaissances nécessite donc cette première interaction entre complexité computationnelle et complexité informationnelle. Le lien entre complexité informationnelle et émergence est mobilisé si on considère l'établissement de connaissances comme un processus morphogénétique. Il est montré en 5.1 que le lien entre forme et fonction est fondamental en psychologie : nous pouvons l'interpréter comme un lien entre information et sens, puisque la sé-

50 Le démon de Maxwell est plus qu'une construction intellectuelle : [Cottet et al., 2017] implémente un démon expérimentalement au niveau quantique.

mantique d'un objet cognitif ne peut se passer d'une fonction. HOFSTADER rappelle dans [HOFSTADTER, 1980] l'importance des symboles à différents niveaux pour l'émergence d'une pensée, qui consistent à un niveau intermédiaire en des signaux. Enfin, la dernière relation entre complexité computationnelle et émergence est celle qui nous permet d'affirmer qu'on s'intéresse particulièrement à une production de connaissance sur des systèmes complexes, les deux premiers pouvant s'appliquer à tout type de connaissance.

Ainsi, toute *connaissance du complexe* embrasse non seulement toutes les complexités et leur relations dans son contenu, mais aussi dans sa nature comme nous venons de montrer. La structure de la connaissance en termes de complexité est analogue à la structure des systèmes qu'elle étudie. Nous postulons que cette correspondance structurelle implique une certaine récursivité, et donc un certain niveau de *réflexivité* (au sens de connaissance d'elle-même et de ses propres conditions).

On peut tenter d'étendre à la réflexivité en tant que réflexion sur le positionnement disciplinaire : suivant [PUMAIN, 2005], la complexité d'une approche est également liée à la diversité des points de vue nécessaires pour la construire. Pour atteindre ce nouveau type de complexité[51], qui serait une dimension supplémentaire liée à la connaissance des systèmes complexes, la réflexivité doit être au coeur de la démarche. [READ, LANE et LEEUW, 2009] rappellent que l'innovation a été rendue possible quand les sociétés ont été capables de produire et diffuser de l'information sur leur propre structure, c'est-à-dire quand elles ont pu atteindre un certain niveau de réflexivité. La *connaissance du complexe* serait donc le produit et le support de sa propre évolution grâce à la réflexivité qui a joué un rôle fondamental dans l'évolution du système cognitif : on pourrait ainsi suggérer de rassembler ces considérations, comme proposé par PUMAIN, sous une nouvelle notion épistémologique de *rationalité évolutive*.

Pour conclure, notons qu'étant donné la loi de la *requisite complexity*, proposée par [GERSHENSON, 2015] comme extension de la *requisite variety* [ASHBY, 1991][52], la *connaissance du complexe* devra nécessairement être *connaissance complexe*. Cet autre point de vue renforce la nécessité

51 Pour laquelle des liens avec les types précédents apparaissent naturellement : par exemple, [GELL-MANN, 1995] considère la complexité effective comme le *Contenu d'Information Algorithmique* (proche de la complexité de Kolmogorov) d'un Système Complexe Adaptatif *observant un autre* Système Complexe Adaptatif, ce qui donne son importance aux complexités informationnelle et computationnelle et suggère l'importance du point de vue d'observation, et par extension de la combinaison de ceux-ci - ce qui est par ailleurs à mettre en relation avec l'approche perspectiviste des sciences complexes présentée précédemment.

52 L'un des principes cruciaux de la cybernétique, la *requisite variety*, postule que pour contrôler un système ayant un certain nombre d'états, le contrôleur doit avoir au moins autant d'états. GERSHENSON propose une extension conceptuelle à la complexité, qui peut être justifiée par exemple par [ALLEN, STACEY et BAR-YAM, 2017] qui introduisent la *requisite variety* multi-échelle, démontrant la compatibilité avec une théorie de la complexité basée sur la théorie de l'information.

de la réflexivité, puisque suivant Morin (voir par exemple [Morin, 1991] sur la production de connaissance), la *connaissance de la connaissance* est centrale dans l'établissement d'une pensée complexe.

Conséquences pratiques

Pour conclure cette section épistémologique, nous proposons de synthétiser l'ensemble des idées introduites sous forme de manifestations concrètes en découlant directement, et qui conditionneront fortement l'ensemble de la forme et de la sémantique de la connaissance introduite par la suite. Ces directions (que nous n'irons pas jusqu'à nommer principes car seulement à l'état d'ébauche) peuvent être regroupées en trois grandes familles : pratiques de modélisation, pratique de la science ouverte, et épistémologie. Sur le plan des pratiques de modélisation, dans chaque section se dégagent différents axes plus ou moins complémentaires :

- La modélisation, qui sera dans la majorité des cas équivalente à la simulation, doit être comprise comme un instrument de connaissance indirect sur des processus au sein d'un système complexe ou sur la structure de celui-ci (d'après la sous-section sur "pourquoi modéliser"), et les modèles devront nécessairement être complexes (d'après la réflexion sur les différents types de complexité) au sens qu'il capturent un phénomène d'émergence faible, tout en respectant des exigences de parcimonie.
- L'exploration des modèles est partie intégrante de l'entreprise de modélisation (voir reproductibilité), et le calcul intensif est un élément clé pour explorer efficacement les modèles de simulation (voir calcul intensif). Les méthodes d'analyse de sensibilité doivent être questionnées et étendues si besoin (comme l'illustre l'exemple de la sensibilité à l'espace).
- Comme suggéré par le positionnement perspectiviste, le couplage de modèles devra jouer un rôle crucial dans la capture de la complexité.

Pour la science ouverte, on peut extraire les points suivants :

- La nécessité de l'ensemble des démarches liées à la science ouverte pour parvenir à la construction de modèles toujours plus complexes, vers la co-construction de modèles par différentes disciplines.
- Dans ce cadre, l'ouverture complète du code source, ainsi que sa lisibilité sont cruciaux. L'explicitation complète du modèle dans le compte-rendu scientifique, ainsi qu'une documentation du code auto-suffisante, sont deux aspects de celle-ci.

- La question des données ouvertes n'est pas négociable dans ce cadre. La quasi-totalité de nos traitements est basée sur des données initialement ouvertes, et lorsque ce n'est pas le cas nous travaillons à un niveau agrégé auquel on peut fournir les données. Les jeux de données construits sont ouverts.

- Concernant les méthodes d'exploration interactive, qui sont un pendant de l'ouverture de la science, nous en développons un certain nombre, mais restons limités par rapport au pré-requis idéal qui devrait rendre celles-ci totalement compatibles avec une démarche reproductible.

Enfin, sur le point épistémologique, on peut également tirer des implications "pratiques" qui seront bien évidemment plus implicites dans notre démarche, mais pas moins structurantes :

- Notre inspiration sera essentiellement interdisciplinaire et cherchera à croiser les différents points de vue.

- Les différents domaines de connaissance (notion que nous préciserons en 8.3, mais qu'on peut comprendre pour l'instant au sens des domaines théorique, empirique et de la modélisation introduits par [LIVET et al., 2010]) sont indissociables pour toute démarche de production scientifique, et nous les mobiliserons de manière fortement dépendante.

- Notre démarche devra comprendre un certain niveau de réflexivité.

- La construction d'une connaissance complexe ([MORIN, 1991]) est ni inductive ni déductive, mais constructive dans l'idée d'une morphogenèse de la connaissance : il peut par exemple être délicat d'identifier clairement des "verrous scientifiques" précis puisque cette métaphore suppose qu'il faut débloquer un problème déjà construit, et de même de faire rentrer notions, concepts, objet ou modèles dans des cadres analytiques stricts, en les catégorisant selon une classification fixe, alors que l'enjeu est de comprendre si la construction des catégories est pertinente. Le faire a posteriori relève d'une négation de la circularité et de la récursivité de la production de connaissance. L'élaboration de modes de compte-rendu rendant compte du caractère diachronique et des propriétés évolutives de celle-ci est un problème ouvert.

⋆ ⋆

⋆

CONCLUSION DU CHAPITRE

La lecture d'un article ou d'un ouvrage est toujours bien plus éclairante lorsqu'on connait personnellement l'auteur, d'une part car on peut profiter des *private joke* et extrapoler certains développements des narrations qui se doivent synthétiques (même si l'art de l'écriture est justement d'essayer de transmettre la majorité de ces éléments, l'ambiance en quelque sorte), et d'autre part car la personnalité a des implications complexes sur la manière d'appréhender la nature de la connaissance et une certaine structure a priori du monde. Pour cela, la connaissance scientifique serait très probablement moins riche si elle était produite par des machines aux capacités cognitives équivalentes, aux connaissances et experiences empiriques subjectives équivalentes et aussi diverses que celles humaines, mais qui auraient été programmées pour minimiser l'impact de leur personnalité et de leur convictions sur l'écriture et la communication (toujours en supposant qu'elles aient une certaine forme de données et fonctions plus ou moins équivalentes). Dans ces laboratoires de recherche dignes de *Blade Runner*, nous doutons que la production d'une connaissance du complexe serait effectivement possible, puisqu'il manquerait à ces machines justement la *rationalité évolutive* développée en 3.3, et nous doutons fortement que celle-ci puisse être produite du moins dans l'état des connaissances actuelles en intelligence artificielle.

Le but de ce chapitre était donc "de faire connaissance" sur les points de positionnements incontournables pour l'ensemble de notre réflexion. Ceux-ci en sont d'autant plus cruciaux car conditionnent très fortement certaines directions de recherche.

Notre positionnement sur la reproductibilité développé en 3.2 implique certains choix de modélisation, notamment l'utilisation univoque de plateformes ouvertes, de workflow et d'implémentations ouverts ; il implique aussi un choix de données qui se doivent au maximum d'être accessibles ou rendues accessibles, et donc certains choix d'objets et d'ontologie, ou plutôt le non-choix de certains : nos problématiques pourraient être mobilisées sur des données d'entreprise fines tout en gardant une cohérence avec l'approche théorique et thématique (la théorie évolutive des villes a largement mobilisé ce type d'étude comme par exemple [Paulus, 2004]), mais la relative fermeture de ce type de données ne les rend pas utilisables dans notre démarche.

Ensuite, notre positionnement sur le rôle du calcul intensif et les besoins d'exploration des modèles 3.1 est source de l'ensemble des expériences numériques et des méthodologies utilisées ou développées.

Enfin, notre positionnement épistémologique 3.3 percole dans l'ensemble de notre travail, et permet de poser les premières briques pour

des formalisations théoriques plus systématiques qui seront développées en chapitre 8.